Calibration of a probabilistic model of DNA evolution

Master's Thesis submitted

to

Prof. Ostap Okhrin

Humboldt-Universität zu Berlin
School of Business and Economics
Institute for Statistics and Econometrics
Ladislaus von Bortkiewicz Chair of Statistics

and

Hugues Roest Crollius, PhD

Ecole normale supérieure
Institut de biologie de l'ENS
Dyogen Group

by

Lucas Tittmann
(533093)

in partial fulfillment of the requirements
for the degree of
Master of Statistics
Berlin, August 16, 2015

Acknowledgement

I would like to thank the group of Hugues Roest Crollius for the discussions which helped to shape this thesis. Especially, I want to thank my supervisor Mr Roest Crollius for all the advice and time which invaluably broadened my understanding of genetics in general, and evolution in particular. Furthermore, I want to thank J. Lucas for the productive team work which helped to improve my programmes and free it from bugs.

Finally, I want to thank Prof. Ostap Okhrin whose open-mindedness to distant applications of statistical methods allowed me to write this thesis in collaboration with ENS Paris.

Abstract

This thesis has two main results: it describes a model of evolution where the DNA is represented by genes, and it describes how optimal parameters for this model can be found. The main focus lies on the estimation of the number of chromosomal events, though gene events are included as a noise factor.

Different methods to statistically estimate the parameters of the model are compared. The adapted estimation methods are applied and estimates for reciprocal translocation and inversion numbers on a phylogenetic tree of 21 Amniota species are provided.

To test theoretical results and calculate error margins, the model was implemented in the gene order simulation software MagSimus, created by the group of H. Roest Crollius at the ENS Paris. Together with an implemented optimization framework, the genome analysing software PhylDiag and the chromosomal event estimation software ChromEvol 2, numerical estimates of the model are calculated for a sub-sample of 5 species. Afterwards, the quality of the simulated genomes is assessed.

Besides the interest in reliable estimates in historic mutation rates alone, the goal of a realistic simulation is the benchmarking of genome order reconstruction programmes. The data was taken from from the Ensembl genome project (Cunningham et al. (2015)), the Genome Size database (Gregory (2015)) and the Time Tree database (Hedges et al. (2015)).

key words: DNA evolution, chromosomal rearrangement, probabilistic model, evolution simulation, estimation

Contents

List of Abbreviations

CDF	Cumulative density function
CI	Confidence interval
DNA	Deoxyribonucleic acid
DCJ	Double-Cut-And-Join
E.g.	For example
etc.	Et cetera
Fig.	Figure
kb	Kilo base pairs = 10^3 base pairs
KS	Kolmogorov - Smirnov
LM	Linear model
M2006	Estimation method based on Mazowita (2006), equation (9)
mb	Mega base pairs = 10^6 base pairs
MRCA	Most recent common ancestor
MS	MagSimus
mya	million years
NNLS	Non-negative least squares
OI	Optimal input
PDF	Probability density function
Q	Quantlet name
SB	synteny block

List of Figures

List of Tables

1 Introduction

One of the major scientific breakthroughs in the second half of the 20th century was the understanding of the genetic code, and ultimately the sequencing of the human genome. At the beginning of the 21st century, modern technology like next-generation sequencing makes it affordable to analyse the genome of hundreds of different species, thus giving us profound insights in the biochemical mechanisms of life. One particular interesting question concerns the development of diversity of life forms, and related to it, the course of evolution of our genetic heritage.

This thesis has two goals related to these questions. Firstly, create and quantify a probabilistic model of gene order evolution. Secondly, implement this model in MagSimus, a simulator for gene order evolution created by the group of H. Roest Crollius to benchmark programmes which try to reconstruct ancestral gene order.

Creating a theoretic model of gene order evolution

At first, we want to create a realistic model of gene order evolution. Genetic data is very complex, and reducing this complexity while keeping important properties which can be analysed statistically is challenging. In the last two decades, various models have been proposed, and based on these models it was tried to estimate various mutation rates, from small nucleotide mutations to large chromosomal rearrangements. Some models are based on small parts of DNA over a short amount of time, e.g. models concerning nucleotide changes, and are therefore even verifiable in test tube experiments. Other models, concerning big scale mutations like chromosomal rearrangements, need high quality assembled chromosomes of many species to draw conclusions over events that happened in the distant past. Therefore, it is only recently that these mutations can be modelled sensibly, and estimations based on these models are still rare and rather rough.

In this thesis, we focus on these large mutations. Confronted with biological reality, the first major part of this work was to screen the literature for available models. One of the earliest approaches proposed was Pevzner and Tesler (2003a), in which they search for so-called breakpoints in sequenced DNA, and try to establish a rearrangement distance between two genomes, i.e. they try to calculate the shortest path to transform one genome into the other, using only inversions and reciprocal translocations. One drawback of this model is that it does not include other mutational events, e.g. gene events which represent a large source of noise. Furthermore, they constrain their model to consider only estimations which can be explained by an exact evolution history, which sometimes leads to features which are biologically hard

to explain. Another popular distance-based approach is based on the double-cut-and-join model (DCJ) (Chauve et al. (2013), pp. 63 - 81). The DCJ is a mathematical elegant way to describe all well-studied rearrangements, but it has the opposite drawback: it also describes rearrangements which are thought not to exist in reality. Therefore, trying to distinguish between purely theoretic model artefacts and real world insights may be difficult. Moreover, although calculating the DCJ distance is done in linear time, finding a scenario which results in the observation is NP-hard (Nelson and Vialette (2008), p. 158). One approach to solve the problem was presented by Miklós and Tannier (2010). They implemented a Markov-Chain-Monte-Carlo method which at first allows all DCJ operations, but then tries to exclude unrealistic scenarios by a temperature based method. Another approach based on the DCJ model is proposed by Lin et al. (2010), which includes gene events but stays rather unspecific on how to use on real data. Zhao and Bourque (2009) proposed another programme, called EMRAE, which infers chromosomal rearrangements. However, they include an operation called transposition which existence in reality is disputed. As this operation can be explained by two, cytogenetically well explained reciprocal translocations, therefore their estimates for the latter are very small. Finally, another approach was described in Sankoff and Mazowita (2005) and Mazowita et al. (2006). They use a model which includes fusions, fissions, reciprocal translocations and inversions, and developed an estimator based on the number of breakpoints in the genome and DNA chunk exchange between chromosomes. As this this model met many criteria we imposed, we decided to follow their approach. However, comparison with real biological data showed that we had to adapt their model in three ways:

1. We change their way to model translocations and inversions.

2. They present an estimator to estimate rates of chromosomal evolution based on their model. However, they do not demonstrate robustness of this estimator under the influence of noise. To this purpose, we compare the results of the estimator on real and simulated data with noise. We show that this noise leads to wrong statistical estimates. Furthermore, we develop an optimization framework for MagSimus to measure the error of their estimation method due to the noisiness of the data.

3. We modify their estimator to better fit our model and evaluate if this adaptation improves the accuracy of their estimator.

4. We show that their model can be used with genomes modelled as a chain of genes instead of a chain of nucleotides (see chapter 7.2.1 for a discussion). We show that despite using

simplified genomes, results are comparable to previous estimates (see chapter 7.1).

The modifications of their model together with justifications for these modifications can be found in chapter 2.4.

Implementing the model in a simulator

The second goal of this thesis is calibrating the simulation MagSimus to create a benchmark test for gene order reconstruction software. It is not possible to directly observe ancient DNA, but we hope that reconstructing the DNA of certain ancestors, e.g. the common ancestor of reptiles and mammals, may lead to new insights in the nature of evolution. There are several different approaches to the problem of DNA reconstruction. The most straightforward method is to try to reconstruct DNA on a nucleobase level, which was done by Paten et al. (2008). This approach, however, does not help understand the inner-workings of the ancestral animals body, as it is difficult to identify genes and other functional elements in this hypothetical DNA. Another approach uses directly the hypothetical gene content of ancestral DNA. The Ensembl project (Cunningham et al. (2015)), amongst other sources, computes and synthesizes the evolutionary history of genes in so-called gene trees. However, not only the gene content, but often also the gene order proofs to be crucial. A gene in a wrong position may lead to a protein not being expressed at all, or at a wrong time point during development, thus being not effective at all. Therefore, not only gene content but also gene order must be inferred. Three programmes which try the latter are MGRA (Alekseyev and Pevzner (2009)), DeCo (Bérard et al. (2012)) and AGORA (Muffato (2010)). For example, AGORA is already used in scientific publications like Berthelot et al. (2015), but a formal benchmarking to assess the quality of its reconstructed gene order is still missing. This would help to decide if particularities found in the reconstructed genome are scientific findings or merely software artefacts. Therefore, a gene order evolution simulator, called MagSimus, was created by the group of Roest Crollius, including the author of this thesis, which was another part of this thesis. Finally, as discussed in chapter 5, the estimation method proposed by Mazowita et al. (2006) did not provide realistic model specifications for our simulation. Therefore, we created an optimization framework in Python to find the optimal input parameters for our simulation. Even more, these numerically found parameters can be seen as numerical estimates of our model. This numerical assessment was the third major part of the work. Theoretically, this framework can be used to find the optimal configuration for all model aspects presented in chapter 2.4. However, the number of variables in combination with a simulation time of MagSimus of several minutes forced us to use this numerical

optimisation only to identify some parameters. For the others, we kept the results from our analysis done in chapter 2.4.

There are already several evolution simulations available, however none can be used to study the aspects of evolution we are interested in. Firstly, there is aevol (Batut et al. (2013)), a bacterial DNA simulation which allows the DNA to be put under different external stress factors and to study their effects on genetic features present in the population. Furthermore, there is ALF (Dalquen et al. (2012)), which tries something similar to MagSimus but again focuses on nucleotide modelling, and hence is not appropriate for gene order analysis.

In this thesis, we focus on a clade of animals called Amniota, which includes mammals, birds and reptiles. They all share one common ancestor about 326 million years in the past, that will be called Amniota for shortness in the rest of the thesis. We use the following approach to accomplish the goals of this thesis:

1. Analysis of available biological data. Reducing complexity of the data by formalizing only interesting aspects of the data (chapter 2.1).

2. Literature analysis to find models transforming known biological mutations into formal operations of our data.

3. Choice of one adapted model. As some of its hypothesis were not fulfilled in our data, we adapted the model in several aspects (chapter 2).

4. Several quantitative parts of the model had to be estimated. Several different methods and programmes which had to be used in order to do so are discussed in chapter 3.

5. Acquire the necessary data (chapter 4).

6. Estimate the quantitative parts of our model with the real data (goal 1, chapter 5).

7. Implement the model in the gene order evolution simulation MagSimus (chapter 3.4). Verify the reliability of the estimates as follows:

   (a) Simulate genomes using MagSimus with estimates from real data.

   (b) Re-estimate the quantitative parts of our model from the simulated genomes.

   (c) Compare.

8. As estimations on real data and simulated data are different, we conclude that our estimation method is not appropriate.

9. We create a numerical optimization framework for MagSimus in order to optimize its parameters (chapter 6). The optimal parameters of MagSimus are at the same time a numerical estimate of the mutation rates. A new score is proposed to measure the quality of our simulated data (chapter 6.3).

10. At the end, we compare our results to the existing literature (chapter 7.1), discuss possible error sources (chapter 7.2), and indicate possible next steps (chapter 8)

As a result of this work, we established a theoretical model of gene order evolution with different types of mutations, including both chromosomal rearrangements as well as gene events. Furthermore, we established an estimation pipeline to infer chromosomal rearrangement rates, namely inversions and reciprocal translocations, for 21 Amniota species. We tested the accuracy of these estimates by selecting 5 Amniota species to use in a software implementation of our model, called MagSimus. We created and used an optimization framework to get numerical estimates of the chromosomal rearrangement rates. We identified the error sources leading to difference between statistical estimate and numerical estimate. Furthermore, we proposed a new score to measure the quality of our simulated data. Finally, by using our numerically optimized parameters as input for MagSimus, we created a functional benchmark test for the gene order construction software AGORA.

Code and data for graphics created with R can be found on `http://quantlet.de/` . The Quantlet names are indicated in the captions with a #. The Python code for the numerical optimization framework together with the code for the Matplotlib graphics will be published together with MagSimus.

2 Biological background and formal model

The first major part of this thesis was to transform biological knowledge into a statistically analysable model. As genetic data is very complex, it is difficult to find a model which on one hand is simple enough to analyse quantitatively in a probabilistic model, but on the other hand does not loose to much detail and makes the results meaningless. In the first part of this chapter, we first describe and then formalize the data. Afterwards, in chapter 2.4, we describe how we modelled operations on our formalized data. After a literature analysis and comparison of several models, which represented the first part of our work, we decided to use the model proposed by Sankoff and Mazowita (2005). However, we found that their model and its assumptions were in several aspects not realistic. Therefore, a second major part of thesis was to analyse the assumptions of their model using descriptive statistics and simulations, and to adapt their model where necessary. Adaptations of the model were, amongst others, the inclusion of several genome operations not available in the original model, changes in the modelling of certain chromosomal rearrangements (reciprocal translocations and inversions), and the adaptation of the model to work with our data. Whereas their data is based on DNA expressed in nucleotides, our data is DNA modelled in genes.

2.1 Definition of a genome

All forms of life are expressed forms of genetic information, which is most of the time coded in Deoxyribonucleic acid (DNA). DNA is a double-stranded organic molecule, which resides in the cellular nucleus. The DNA is made of a polymer of nucleotides, each nucleotide consisting of a certain sugars, a phosphate group and four different nucleobases. The four different nucleobases are Adenine (`A`), Cytosine (`C`), Guanine (`G`), and Thymine (`T`). The order of nucleotides are used as letters to encode the genetic information. A typical Amniota genome has between 0.9×10^{109} and 4.1×10^{109} nucleotides. A genetic word consists of three such nucleotides and is called a codon. These codons code for one of 20 possible amino acids. Proteins themselves are besides lipids the primary organic substance in many animals, and are used in nearly all body functions. Because proteins are that important, a particularly interesting part of the DNA is a chain of codons which code for a protein, and this is called a gene. A gene in the human genome can be between 300 and 2.3 $\times 10^6$ bases long. There are about 22,000 coding genes in the DNA, and depending on how the length of genes is measured, they make up from 16% to 43% of human DNA. The rest of the DNA has either regulation functions, unknown or no functions at all. Finally, a gene has an orientation, called

sense or anti-sense.

We can now formally define a gene as g, a tuple of a name n and an orientation o:

$$g = (n, o)$$

$o \in \{0, 1\}$. Furthermore, we define an operation called flip ($\mathcal{F}$), which works as follows: $\mathcal{F}(0) = 1$ and $\mathcal{F}(1) = 0$. We can generalize this definition to be applicable to gene g as $\mathcal{F}(g) = \overline{g} = (n, \mathcal{F}(o))$.

DNA can either be one long polymer, or, as in all Amniota species, be split up in several pieces of polymer. Each of these DNA pieces is called a chromosome. If we ignore the intergenic DNA, we can define a chromosome c in a simplified version as a tuple of genes:

$$c = (g_1, g_2, \ldots, g_n)$$

where n is the number of genes on this chromosome. In chapter 7.2.1, we discuss in detail the assumptions and effects of defining a genome like this. If defined like this, a chromosome is largely defined by the underlying gene order. As a chromosome does not have an orientation, an equivalent representation of chromosome c would be $c = (\overline{g_n}, \overline{g_{n-1}}, \ldots, \overline{g_1})$. The complete set of chromosomes is called a genome G:

$$G = \{c_1, c_2, \ldots, c_m\}$$

where m is the number of chromosomes in the genome. The chromosomes are named by their size, where c_1 indicates the largest chromosome in the genome. The standardized chromosome size of c_j is defined as

$$c_i = \frac{c_j}{\sum_{i=1}^{n} c_i} \tag{1}$$

.

Due to the way genetic inheritance and recombination works, most Amniota species have a pair of unequal sex chromosomes, called allosomes, and some non-sex chromosomes, typically between 8 and 60, called autosomes. The autosomes come in pairs, i.e. one is inherited from the father and one very similar chromosome from the mother. E.g., the human has in total 46 chromosomes, including two allosomes (X and Y) and 44 autosomes, called chromosomes 1 to 22, each in two versions. For simplicity, we only consider one chromosome from each

pair of autosomes. Furthermore, we only include the bigger of the two allosomes, i.e. the X-chromosome in our selected mammals and the Z-chromosome for birds (see chapter chapter 4 for a discussion).

2.2 Definition of genome operations

During cell division, DNA needs to be copied in a process called DNA replication. The DNA in some specific cells called gametes is transferred to the next generation. Yet, copying is not accurate, and many types of errors may occur. This includes single nucleotide polymorphisms, genic events and chromosomal rearrangements. They differ in the size of the involved region (1 nucleobase up to whole chromosome and genome changing mutations) and their frequency (several occurrences per generation up to only one event every hundred million years). Though the rates of frequent occurring errors, like single-nucleotide changes, can be measured in test tube experiments, the large rearrangements are so rare that performing an experiment is pointless. Hence, it is necessary to infer the historic rates of these errors, which is one of the goals of this work.

As discussed in the previous chapter, we define our genomes as an order of genes, hence we do not model single-nucleotide mutations or other mutations which do not affect gene order. This includes chromosomal rearrangements which do not change gene order. We include the following 7 operations in our model, including 3 gene event, i.e. operations which only change one gene, and 4 chromosomal rearrangements, i.e. operation which can include more than one gene. In the following, genome G is defined as

$$\begin{aligned} G &= \{c_1, c_2, \dots\} \\ &= \{(g_{1,1}, g_{1,2}, \dots), (g_{2,1}, g_{2,2}, \dots), \dots\} \\ &= \{((n_{1,1}, o_{1,1}), (n_{1,2}, o_{1,2}), \dots), \dots\} \end{aligned}$$

Furthermore, we call a pair of consecutive genes an *adjacency*. In our representation of the genome, adjacencies represent intergenic regions. We can compare the adjacencies of genome G_1 with genome G_2. Every adjacency present in G_1 but not in G_2 is called a *breakpoint* (compare Gascuel (2007), p. 331). Even more, we specifically require a breakpoint to be caused by a chromosomal rearrangement. A breakpoint according to the former definition which was not caused by a chromosomal rearrangement will be called a pseudo-breakpoint. In our model, pseudo-breakpoints are caused by genic events. Furthermore, we define a

synteny block to be a sequence of consecutive adjacencies between two breakpoints. With our definitions, we can use the following equation to relate the number of synteny blocks s, the number of breakpoints b and the number of chromosomes c_1 and c_2 in genomes G_1 and G_2 (compare Mazowita et al. (2006)):

$$s = b + \max(c_1, c_2) \tag{2}$$

Gene duplication

A gene may be copied, and the duplicate may be inserted somewhere into the genome. The orientation of the duplicate can be different from the original gene. We can describe the duplication $\mathcal{G}_{\mathcal{D}}$ on the genome G of gene $g_{i,j} = (n_{i,j}, o_{i,j})$, where i is the number of the chromosome and j is the number of the gene in chromosome i, as follows:

$$\mathcal{G}_{\mathcal{D}}(G, g_{i,j}) = \{\dots, (\dots, (n_{k,l}, o_{k,l}), (n_{i,j}, o^*_{i,j}), (n_{k,l+1}, o_{k,l+1}) \dots), \dots\}$$

,

where k, l and $o^*_{i,j}$ can, but do not have to differ from i, j and $o_{i,j}$, respectively. A *tandem duplication* is a duplication where the copy is created directly at either side of the original gene, i.e. the number of genes between the duplicates, the *gap*, is 0. A duplication which is not a tandem duplication is called a *distant duplication*.

Gene deletion

A gene may be erased from the genome. We can describe the deletion $\mathcal{G}_{\mathcal{E}}$ on the genome G of gene $g_{i,j}$, where i is the number of the chromosome and j is the number of the gene in chromosome i, as follows:

$$\mathcal{G}_{\mathcal{E}}(G, g_{i,j}) = \{\dots, (\dots, g_{i,j-1}, g_{i,j+1}, \dots), \dots\}$$

.

Gene birth

A gene may be appear in the genome. We can describe the birth $\mathcal{G}_{\mathcal{B}}$ on the genome G of gene $g_{i,l+1}$, where i is the number of the chromosome and l is the number of genes in chromosome i, as follows:

$$\mathcal{G}_{\mathcal{B}}(G, g_{i,l+1}) = \{\dots, (\dots, g_{i,j}, g_{i,l+1}, g_{i,j+1}, \dots), \dots\}$$

.

Fission

During a chromosome fission $\mathcal{C}_{\mathcal{S}}$, a chromosome c_i of genome G is split up in two parts:

$$\mathcal{C}_{\mathcal{S}}(G, c_i) = \{\dots, (\dots, g_{i,j}), (g_{i,j+1}, \dots), \dots\}$$

.

The splitting of the chromosome at adjacency $(i, j) - (i, j + 1)$ creates a breakpoint.

Fusion

During a chromosome fusion $\mathcal{C}_{\mathcal{F}}$, two chromosomes c_i and c_k of genome G are fused at one of their extremities. As every chromosome has 2 extremities, there are 4 possible scenarios:

$$\mathcal{C}_{\mathcal{F}}(G, c_i, c_j) = \begin{cases} \{\dots, (\dots, g_{i,j-1}, g_{i,j}, g_{k,1}, g_{k,2} \dots), \dots\} \\ \{\dots, (\dots, g_{i,j-1}, g_{i,j}, \overline{g_{k,l}}, \overline{g_{k,l-1}} \dots), \dots\} \\ \{\dots, (\dots, \overline{g_{i,2}}, \overline{g_{i,1}}, g_{k,1}, g_{k,2} \dots), \dots\} \\ \{\dots, (\dots, \overline{g_{i,2}}, \overline{g_{i,1}}, \overline{g_{k,l}}, \overline{g_{k,l-1}} \dots), \dots\} \end{cases}$$

.

It is possible, though unlikely, that a fusion destroys a breakpoint created by a fission. This happens if two chromosomes which were split by a fission are joined at the exact same extremities which were involved in the fission.

Reciprocal translocation

During a reciprocal translocation $\mathcal{C}_{\mathcal{T}}$, two chromosomes c_i and c_k of genome G exchange two non-empty extremities. In the formal description, though not in cytogenetic reality, a reciprocal translocation can be seen as a series of 2 fissions and 2 fusions, thus creating 2 breakpoints. The two breakpoints in this example are at positions $(i, j) - (i, j + 1)$ and $(k, l) - (k, l + 1)$. There are 2 different scenarios:

$$\mathcal{C}_{\mathcal{T}}(G, c_i, c_j) = \begin{cases} \{\dots, (\dots, g_{i,j-1}, g_{i,j}, g_{k,l+1}, g_{k,l+2} \dots), (\dots, g_{k,l-1}, g_{k,l}, g_{i,j+1}, g_{i,j+2} \dots), \dots\} \\ \{\dots, (\dots, g_{i,j-1}, g_{i,j}, \overline{g_{k,l}}, \overline{g_{k,l-1}} \dots), (\dots, \overline{g_{k,l+2}}, \overline{g_{k,l+1}}, g_{i,j+1}, g_{i,j+2} \dots), \dots\} \end{cases}$$

.

Inversion

During a chromosome inversion $\mathcal{C}_{\mathcal{I}}$, a part of chromosome c_i of genome G is inverted. As in a translocation, in the formal description, though not in cytogenetic reality, an inversion

can be seen as a series of 2 fissions and 2 fusions, thus creating 2 breakpoints. The two breakpoints in this example are at positions $(i, j) - (i, j + 1)$ and $(i, l) - (i, l + 1)$.

$$\mathcal{C}_{\mathcal{I}}(G, c_i) = \{\dots, (\dots, g_{i,j}, \overline{g_{i,l}}, \dots, \overline{g_{i,j+1}}, g_{i,l+1} \dots) \dots\}$$

.

2.3 Phylogenetic trees

We define a phylogenetic tree as a connected, acyclic graph with *edges* (also called *branches*), *vertices* (also called *nodes*) and a root (Gascuel (2007), pp. 3-4). We can define the degree of a node n as the number of branches containing n (ibidem). Nodes with a degree of 1 are called a *leaf* or *modern genome*, nodes with degree of 3 and more are called *internal nodes* or *ancestral genomes* (ibidem). Additionally, there is one node with a degree of 2, which is called the root (ibidem). If all internal nodes have a degree of 3, the tree is called fully resolved (ibidem). Each edge has a non-negative length which may be measured in years, number of breakpoints, number of chromosomal events, or any other distance measure. A *path* is a set of edges, and the length of a path is defined as the sum of all edge lengths on the path. Furthermore, we define the distance between two modern genomes as the length of the shortest path connecting them. More specifically, the *phylogenetic distance* is defined as the distance between two genomes measured in years.

We call the internal node C a *common ancestor* of modern genome A and B if C is contained in both, one edge on the path from A to the root and the path from B to the root. The common ancestor which has the shortest path length to either of the modern genome is called the *most recent common ancestor* (MRCA) of the two modern genomes. In our case, the root of the tree is the most recent common ancestor of all Amniota genomes, for short called Amniota. We call all branches by the node contained in the branch with the shortest path length to a modern genome. An example of a phylogenetic tree including all our selected Amniota genomes can be found in appendix, Fig. 15.

2.4 Model design

The following section specifies how we designed our probabilistic model of gene order evolution. In particular, it discusses biological knowledge or studies we used to design the operations in our model.

However, there are four topics which had to be resolved quantitatively using different estimation techniques:

1. The rate and placing of gene events
2. The rate of fusions and fissions
3. The rate of inversions and translocations
4. The size distribution of inversions

The first represents the noise source for our main interests, questions 2-4. Therefore, we are not directly interested in this number, and a rough estimate of it is sufficient. Chapter 2.4.4 discusses how we modeled and calculated gene events.

However, the latter three form an important part of our model, and the estimates themselves are interesting. Therefore, we used more advanced methods, discussed in chapter 5. A summary of our decisions can be found in Table 4.

Furthermore, we took the following design decisions for our model: the number of events on a branch of the phylogenetic tree in our model is fixed, i.e. they are no random variables. The only source of randomness in our model comes from the order of the operations, not from their number. We argue that this additional factor of chance will not provide any insights in the estimation of the unknown quantitative parameters, and instead just add noise. In the future, however, we may relax this conditions to further test the robustness of ancestral genome reconstruction software as AGORA. One possible alternative to fixed event numbers on branches are Poisson distributed random variables with our estimated event number as a mean.

2.4.1 Fusions and Fissions

Fusion and fissions are large chromosomal rearrangements which have a low occurrence rate in Amniota species. In most lineages the rates are considered to be lower than one event per million years. They can be seen as opposing operations, as a fission increases the chromosome count by one and a fusion decreases the chromosome count by one. However, a fission creates a breakpoint, a fusion does not.

To model a fission, three things must be considered: (a) which chromosome is split, (b) how often chromosomes are split, and (c) where the chromosome is to be split. Likewise, for a fusion we ask: (a) which chromosomes are fused, and (b) how often chromosomes are fused.

(a) Which chromosomes are involved

We consider two different possibilities: either the two chromosomes are chosen independently of their size, i.e. each chromosome has the same chance to be part of a fusion, or the chance to be part of the fusion does depend on chromosome size. During a fusion, the ends of two chromosomes get connected. As all chromosomes have exactly two extremities, the size of the chromosomes should not influence the probability to be part of a fusion. However, as discussed in Ferretti et al. (1996), there may be a physical limit to chromosome size, imposed by the size of the nucleus. In our data, however, there does not seem to exist such a barrier for all species (see Fig. 1). The same considerations, but concerning the lower limit, can be done in regard to fissions. Birds have micro-chromosomes with only a few genes, therefore it does not seem convincing to introduce boundaries as proposed by Ferretti et al. (1996). We therefore decided to choose chromosomes for inclusion in a fusion or fission independently of their size. Even more, simulations showed that due to the rare number of fusion and fissions, this choice has only a small effect on the results.

(b) How often is the operation executed

The number of fusions and fissions in the phylogenetic tree: We decided to quantify these parameters based on the number of chromosomes in modern genomes and an estimation programme called ChromEvol 2 (Glick and Mayrose (2014), see chapter 3.1). Our results are presented in chapter 5.2.

(c) Where is a chromosome split

To assess where chromosomes should be split during a fission, we follow the analysis of Pevzner and Tesler (2003c). They argue that there may be fragile regions in the genome, i.e. regions in the genome where higher number of breakpoints are found. However, these regions may be distributed uniformly over the chromosome, hence a uniform distribution of breakpoints may be a valid null hypothesis for genomes with lower resolutions of detail. As our representation of genomes in form of genes is a low-resolution representation, we adapt the model of uniform breakpoint distribution over the chromosome, and therefore chose the point to split the chromosome during a fission uniformly over the chromosome.

2.4.2 Reciprocal Translocations

There are three aspects of reciprocal translocations which have to be modelled: (a) which chromosomes are involved, (b) how large the involved segments should be and (c), how many

there are.

(a) Which chromosomes are involved

The selection mechanism of chromosomes for a reciprocal translocation was subject to several papers. Ferretti et al. (1996) proposes three different models:

1. Uniform chromosome selection: chromosomes are selected independent of their size.
2. Proportional chromosome selection: chromosomes are selected relative to their size.
3. Proportional chromosome selection with boundaries: chromosomes are selected relative to their size, and translocations leading to extreme sized chromosomes are rejected.

They calculate the limit size distribution of chromosomes for 2 and 3 chromosomes and use computer simulations to calculate the limit size distribution for a genome with realistic chromosome numbers. They find that (1), a chromosome selection independent of size, creates more evenly sized chromosomes than (2). Chromosomes are rather evenly sized in reality, therefore model (1) is more appropriate. However, even under model (1), the larger chromosomes are too large and the smaller chromosomes too small compared to real chromosomes. Therefore, they suggest model (3) which sets boundaries for chromosome sizes, and rejects translocations if the resulting chromosomes become too large or small. However, it is difficult to justify such boundaries over a large range of genomes, as shown in Fig. 1.

De et al. (2001) suggest modelling centromeres, i.e. centres of chromosome, which cannot be exchanged freely. Their model slightly improves the chromosome size distribution to be more uniform compared to the proportional model. Furthermore, they provide a formula to calculate the limiting distribution of the proportional model. However, ultimately they also suggest using upper limits on chromosome sizes. As general size limits seem not to exist in our data, we decided to use the uniform chromosome selection.

(b) What are the sizes of chromosome fragments

As suggested by Sankoff and Mazowita (2005), we draw for both chromosomes randomly one position to split it, and choose randomly one fusing scenario as described in chapter 2.2 (choosing the breakpoint at random at lower resolution is suggested both by Sankoff and Mazowita (2005) and Pevzner and Tesler (2003b)). Fig. 16 shows the chromosome size distribution of modern genomes as well as the simulated limit distributions of the uniform and proportional chromosome selection model.

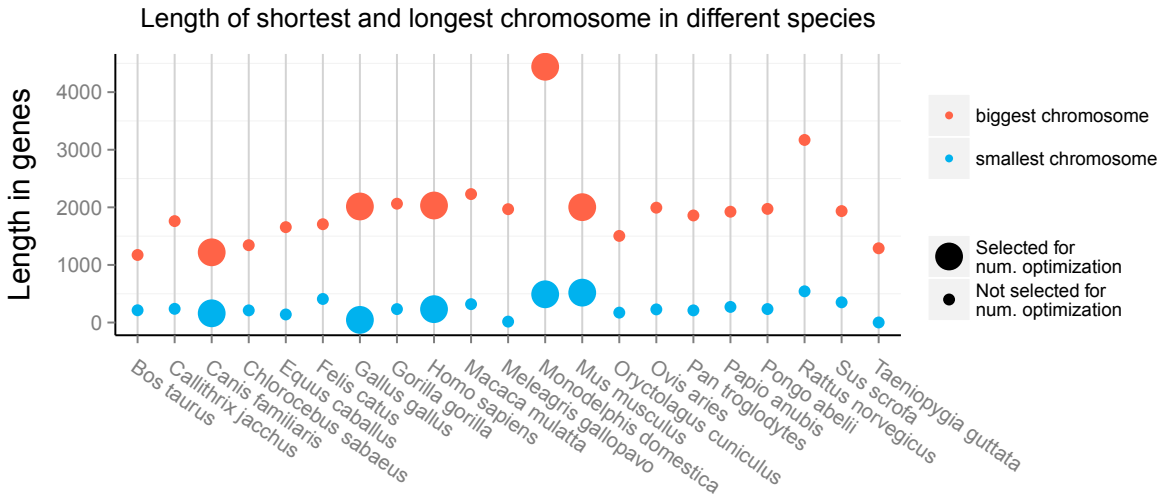


Figure 1: Minimal and maximal chromosome sizes in gene numbers

The upper, red dots indicate the size of the largest chromosome of a species, the lower blue dots indicate the size of the smallest chromosome in a species, not including Y and W chromosomes. The larger dots indicate species which were selected to be simulated with the genome simulator MagSimus. As can be seen, there is large variation in maximal and minimal chromosome size, and hence no conclusion can be drawn on physical constraints on upper or lower chromosome sizes. There are both, very large macro-chromosomes in Monodelphis domestica as well as micro-chromosomes in birds as Gallus gallus (The same holds true for size measured in nucleotides, see appendix, Fig. 16). #ChrSizesMinMaxAmniota

(c) How often is the operation executed

To estimate the number of reciprocal translocations, we follow the approach of Mazowita et al. (2006). The details how they use the number of dispersed chromosome fragments to estimate the number of reciprocal translocations can be found in chapter 3.2.1.

2.4.3 Inversions

There are three aspects of inversions which have to be modelled: (a) where the inversion takes place, (b) how large the inverted segment should be and (c), how many inversions there are. In reality, inversions appear to have varying sizes, thus inversion size is modelled as a random variable.

(a) Which chromosome is involved

Sankoff and Mazowita (2005) suggest that inversion size I, measured in nucleotides, should be modelled as $I = 10^X, X \sim \Gamma(6.539, 0.447)$. This sets the median inversion length at 600 bases. In our model, however, distances on chromosomes are measured in genes. A gene in the human genome has a median size of 24 kb for the longest transcript (mean size = 44 kb).

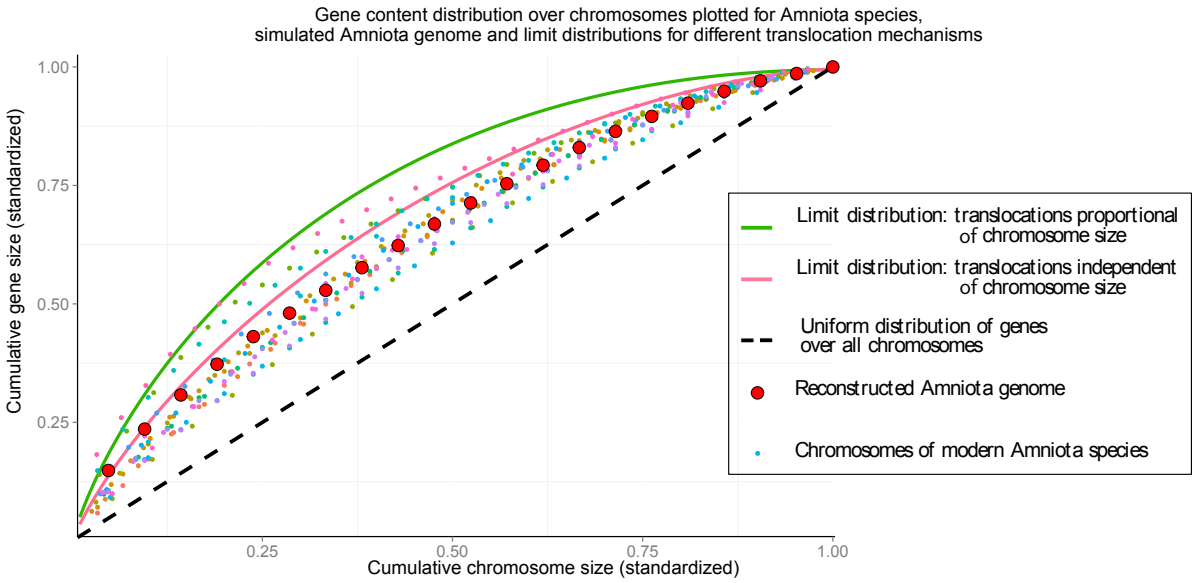


Figure 2: Cumulative density function for selected Amniota species and estimated Amniota start genome

THe abscissa measures genome size in standardised chromosomes, e.g. 1 chromosome of the 23 human chromosome genome stands for $\frac{1}{23} = 4.3\%$, whereas ordinate measures genome size in standardized genes, e.g. 1000 genes of 20,000 are 5%. Chromosomes are sorted by increasing size for each species. The black line indicates the theoretical uniform distribution, i.e. all chromosomes consist of the same number of genes. Furthermore, the simulated limit distribution for uniform and proportional random sampling of translocations are shown. Real genomes are closer to uniform chromosome sizes than the data simulated by our model. #ChrSizesDistributionsInAmniota

Even more, a gene including non-coding DNA around it has a mean size of 150 kb (Fig. 13). Using their distribution, 89% of the inversions will be smaller than one medium sized gene. Therefore, we expect to underestimate the number of inversions. Furthermore, as their distribution placed much emphasis on inversions which cannot be seen with our genome representation, we chose to assess the inversion distribution using a simulation in chapter 6.2. As a result, we model our inversion sizes in genes as $I \sim \Gamma(1, 13.69)$ (see Fig. 23).

When executing an inversion, we at first draw an inversion size I. Afterwards, we randomly select one chromosome c which is at least one gene longer than I. Finally, we randomly chose a position on the chromosome amongst all positions such that I steps added to the position still falls into the chromosome. Sankoff and Mazowita (2005) proposes to draw the position independently of the fact whether the inversion can be executed, and in case that it cannot, to redraw I and the position. However, as it is more likely that big inversions cannot be placed on a certain position, e.g. if the chromosome is smaller than I, this changes our empirical inversion size distribution, and will decrease the number of executed large inver-

sions. To keep the empirical inversion size distribution as close as possible to our theoretical, we only redraw in the case that no position can be found in all the genome to execute that inversion.

(c) How often is the operation executed

Finally, we can use the following equation to infer the number of inversions between two genomes G_1 and G_2, given the number of reciprocal translocations. Be i the number of inversions between both genomes, s the number of synteny blocks, c_1 (c_2) the number of chromosomes in G_1 (G_2), f the number of fissions and t the number of reciprocal translocations between both genomes, then

$$i = \frac{s - \min(c_1, c_2) - f}{2} - t \tag{3}$$

However, this equation only holds if no breakpoints are reused. E.g., if a chromosome is split during a fission at a breakpoint, the number of synteny blocks s will remain constant while the number of fissions f will increase, thus lowering the number of estimated inversions i. As the genomes are large in comparison to the number of chromosomal rearrangements f, t and i, we consider breakpoint reuse in our genomes to be of low importance.

2.4.4 Gene events

Given the number of gene events, it must be decided which gene gets duplicated or deleted. As we do not assign functions to the genes in our genome, we cannot use biological knowledge to determine which genes are more likely to be subject to a gene event. Therefore, we assign each gene the same probability to get duplicated or deleted. This may be debated, as in reality, certain olfactory genes are duplicated several hundred times, whereas other genes most likely never got duplicated since the Amniota genome. To address this weakness in our model, a future version may be directly using all information of the gene trees gene histories, i.e. delete and duplicate genes as indicated in the gene trees. This future model would create one gene family with 900 genes instead of distributing the 900 duplications randomly over all genes as it is done in the present model.

The number of gene events can be directly deduced from the gene trees. Gene trees allow to place a number of gene births, duplications and deletions on each branch in the phylogenetic tree. However, these numbers may be misleading, as they are subject to many possible errors (see chapter 7.2.2). For example, the gene trees may indicate 1,220 duplications and 4,774 gene deletions on the human lineage (see Fig. 6). However, it is possible that there were

1,221 duplications and 4,775 gene deletions, where the additional gene deletion neutralized the additional gene duplication. Hence, these numbers represent the parsimonious approach and indicate only observable events, and most likely underestimate the true numbers. As a consequence, we constrain our model to never delete a gene which got duplicated. For example, we know that in reality we observe 1,220 duplicates in the human genome. As we force our model to execute exactly 1,220 duplications, we cannot delete one of these duplicates, and neither the original gene, otherwise we would only observe 1,219 duplications. Hence, we choose a random, non-duplicated gene for deletion.

Furthermore, we have to decide where to place gene births and gene duplications. In our simplified genomes, there is no information which can help placing a gene birth, hence every position in the genome has the same probability to be chosen for a gene birth. For gene duplication, we use a different approach. It is known that duplications can either happen close to the original gene or distant to it. In order to decide which model is appropriate, we measured the distance between duplicates. Consider the following sequence: AABAB. The only thing we know is that the original sequence was AB. In our approach, we count one duplication with a gap of zero genes (AA), and two duplications with a gap of one gene (ABA and BAB). There are, however, several other possible scenarios to explain the outcome. For example, there could have been two tandem duplications, resulting in the sequence AABB, followed by a duplication of gap size 2, copying left A between the Bs. When using the parsimonious approach, we will underestimate the number of tandem duplications, as groups of tandem duplicates may be parted by distant duplications. As there hundreds of duplications leading to numerous scenarios, we decided to pursue the parsimonious approach despite this possible source of error. Duplicates on other chromosomes cannot be directly assigned a distance, and hence the gap is set to infinite. By looking at the distribution of distances of all duplicates, we can gain insight into the duplication behaviour. However, this direct approach only works in modern species where gene order is known. For ancestral genomes, we only know gene content, and thus which duplicates where present in the genome, but not where. To infer the duplication distances in these genomes, we used the following approach: For every duplicated gene in the ancestral genome, we measured the duplication distance in all modern versions of these genes and took the minimum. For example, there is a duplication of gene G in the Euarchontoglires branch, creating gene G^*. In the human genome, we observe a gap of 5 between G and G^*. In the mouse genome, due to another duplication, we observe two versions of G^*. One has a gap of 1 to the original gene, the other

is on another chromosome. Our assigned duplication distance would be the minimum of all these gaps, and therefore be 1.

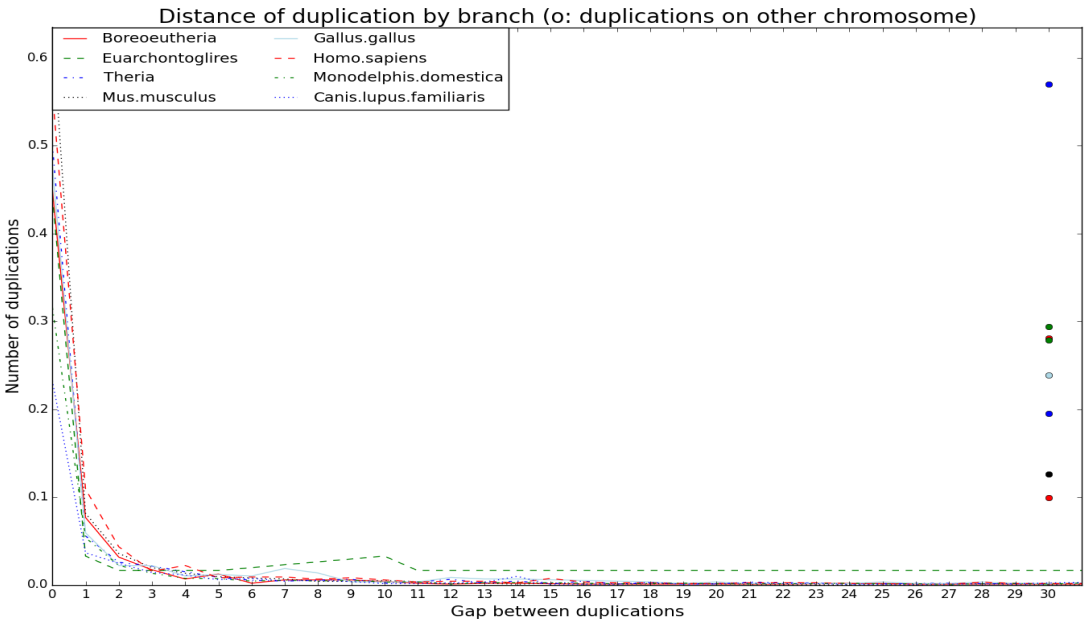


Figure 3: Density functions for duplication distances measured in gaps of genes

Canis lupus has only about 23% tandem duplicates, whereas Mus musculus has more than 58% of its duplications appearing directly at the side of the original gene. In the same notion, nearly 58% of Canis lupus duplications appear on other chromosomes than the original gene (blue dot), whereas this is true only for 13% of the Mus musculus duplications. This shows that the proportion of close to distant duplications varies heavily between different species. However, all distribution have in common that after a rapid decrease, they stay nearly constant after gap 3, indicating the existence of two types of duplications: short distance duplications, where appearance depends heavily on the distance to the original gene, and distant duplications, which appear uniformly over the genome.

Fig. 3 shows our results for the 5 genomes selected for numerical optimization, including their ancestral genomes. The distance of a duplication seems to be build of two parts: either a duplication appears within a range of gap 3, with most of these duplications being tandem duplications, or it appears randomly somewhere in the genome. We argue that most close distance duplications are, in fact, tandem duplications. They are only measured with a gap of 1 or 2 due to other genes, e.g. distant duplicates of other genes, appear between the two duplicates. Hence, we model the duplication distance in two steps: first, a Bernoulli random variable decides with probability p if a duplicate is a tandem duplicate. If it is a tandem duplicate, the duplicate appears with equal probability to either side of the original gene. If it is not a tandem duplicate, all positions of genome have the same probability to be drawn for the placement of the duplicate. Finally, as a result of an analysis done by Joseph Lucas,

in 75% of duplications we set the orientation of the duplicate to be equal to the original gene, and in 25% we inverse it.

Furthermore, Fig. 3 indicates that there is a large variety of tandem duplication proportions. This was a surprising result, and we were interested if this holds true for other Amniota species. Fig. 17 shows that there is a large variety in all our selected Amniota species, which is an interesting research topic in future work. As a direct implication for the model, we decided to make the tandem duplication proportion branch specific.

3 Methods

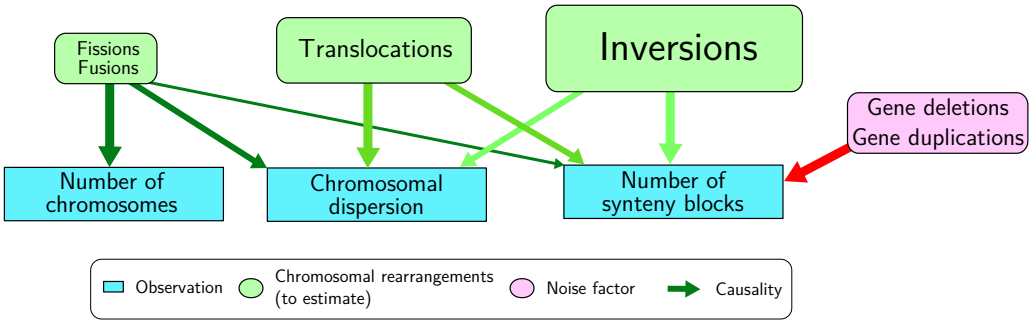


Figure 4: Overview of causality structure of chromosomal rearrangement rates (green, to be estimated) and observations (blue)

The thickness of lines indicate strength of dependency, the sizes of the boxes indicate the absolute occurrence of events. As inputs have effects on several outputs, estimation becomes difficult, especially in presence of the common noise factor, gene events.

This chapter describes the methods used to estimate the quantitative parameters of the model presented in the previous chapter. These parameters include

1. The number of fusions and fissions on each branch of the phylogenetic tree
2. The number of reciprocal translocations and inversion on each branch of the phylogenetic tree
3. The size distribution of the inversions

As shown in Fig. 4, there are several independent variables to estimate the number of fusions and fissions. However, only one variable, the number of chromosomes, allows an estimation without an important noise factor. Therefore, we used an estimator based on a continuous time Markov process of chromosome numbers, implemented in the software ChromEvol 2 (chapter 3.1).

To estimate the number of reciprocal translocation per branch, we at first calculate the translocation distance between two genomes using a measure of chromosomal dispersion (chapter 3.2). The inversion distance between two genomes can be estimated with another independent variable, the number of synteny blocks, calculated with the software PhylDiag (Lucas et al. (2014)). PhylDiag helps to filter out the important noise, gene events, and thus allows us to accurately measure the number of synteny blocks. Using equation 3 and plugging in the previous estimates, we can then infer the the inversion distance. Finally, we use non-negative least squares estimation to estimate the branch lengths of the tree (chapter 3.3). A schematic representation of this approach can be found in Fig. 6.

Finally, the size distribution of the inversions was inferred by computer simulations (chapter 6).

3.1 ChromEvol

ChromEvol 2 is a estimation and simulation software to infer chromosome numbers of ancestral species created by Glick and Mayrose (2014). It is based on a continuous time Markov process. The following paragraphs paraphrase the explication of Mayrose et al. (2010), describing the first version of ChromEvol.

In the most basic form, the model uses a instantaneous rate matrix Q defined as

$$Q_{i,j} = \begin{cases} \lambda & j = i+1 \\ \delta & j = i-1 \\ 0 & \text{otherwise} \end{cases}$$

for $i \neq j$, where λ is the rate of chromosome gains, i.e. the fission rate, and δ is the rate of chromosome losses, i.e. the fusion rate. The parameter i is the old number of chromosomes, which either increases or decreases by 1 to the new number of chromosomes j. Each row should sum to zero, therefore the diagonal elements must be (Mayrose et al. (2010), p. 133):

$$Q_{i,i} = -\sum_{i \neq j} Q_{i,j}.$$

This corresponds to the M0 model of Mayrose et al. (2010). The additional operations of the other models are not observed in Amniota species, hence they can be ignored. For easier computation, they only allow for integer chromosome numbers between 1 and C, where C is x chromosomes bigger than the largest observed chromosome number. For large enough x and small transition probabilities, this truncation will only marginally skew the distribution.

The probability $P_{i,j}$ to get from state i to state j over a time t can be calculated as (Mayrose et al. (2010), equation 3)

$$P_{i,j}(t) = e^{Qt} = \sum_{m=0}^{\infty} \frac{(Qt)^m}{m!}$$

If we assume that events happen independently on different branches, it is now possible to calculate the likelihood of the tree. Let π_i be the probability of the root R having i chromosomes, and $P(D|R = i, \lambda, \delta)$ be the probability of observing data D if the root has i chromosomes and known rates λ and δ, then the likelihood of the tree can be calculated as (Mayrose et al. (2010), equation 5)

$$L = \sum_{i=1}^{C} \pi_i P(D|R = i, \lambda, \delta)$$

with π_i being set to the relative probability to see D with i chromosomes at the root compared to all other chromosome numbers at the root, i.e. $\pi_i = \frac{P(D|R=i,\lambda,\delta)}{\sum_{j=1}^{C} P(D|R=j,\lambda,\delta)}$.

Using this likelihood, the Akaike Information Criterion (AIC) can be calculated. However, the calculation of the likelihood is computationally expensive to calculate if the tree is large. Therefore, Mayrose et al. (2010) use the pruning algorithm proposed by Felsenstein (2004), pp. 251-255. It is a recursive way of calculating the expression from above, starting at the leafs of the tree. Finally, they use numerically optimization to find the maximum likelihood estimates for the rates λ and δ. In our analysis, we also consider the special case $\lambda = \delta$, which can be calculated in the same way.

3.2 Estimating reciprocal translocation and inversion distance

As discussed in previous chapters, there are different models which allow to reduce complex genome data in order to estimate certain properties of evolution like mutation rates. Mazowita et al. (2006) proposed an estimator to infer the reciprocal translocation and inversion distance between two genomes. Furthermore, they suggested a pipeline to reduce DNA expressed in nucleotides into two statistics which can be used as input to that estimator. In chapter 3.2.1, we explain the estimator of Mazowita et al. (2006). Furthermore, we introduce a variant of their estimator which is more adapted to our model. In chapter 3.2.2, we explain how we calculate the input statistics for the estimator based on our DNA expressed in genes.

3.2.1 Chromosomal rearrangement estimation according to Mazowita (2006)

Mazowita et al. (2006) propose an estimator for the number of reciprocal translocations using the number of dispersed chromosome fragments in the chromosome, and the following paragraphs are a summary of Mazowita et al. (2006) with changed variable names.

Let G_1 and G_2 be two genomes with m and n chromosomes. Let $c_{1,1}, \ldots, c_{1,m}$ be the chromosomes of genome G_1, and $c_{2,1}, \ldots, c_{2,n}$ be the chromosomes of genome G_2, and their standardised chromosome sizes be $p_{1,1}, \ldots, p_{1,m}$ and $p_{2,1}, \ldots, p_{2,n}$ as defined in equation 1. A reciprocal translocation is executed as follows: At first, a chromosome $c_{1,i}$ is chosen from genome 1 with $p_{1,i}$. Afterwards, a second chromosome $c_{1,j}$ is chosen without replacing the first chromosome, hence the probability for the second chromosome to be selected is $\frac{p_{1,j}}{1-p_{1,i}}$. After t translocations in total, $c_{1,i}$ is therefore included in approximately $2tp_{1,i}$ (Mazowita et al. (2006), equation 5).

We can see G_2 as the result of many reciprocal translocations executed on G_1. Hence, every chromosome in G_2 corresponds to one original chromosome in G_1. Furthermore, the probability of $c_{2,j}$ not being included in a translocation is $p_{2,j}$, hence the probability for it to be not included in a translocation after t translocations is therefore $(1 - p_{2,j})^{2t}$. Hence, the probability of $c_{1,i}$ and $c_{2,j}$ being never in a translocation is $(1 - p_{2,j})^{2tp_{1,i}}$. We can now sum over all chromosomes in G_2 to get $v_{1,i}$, the expected number of chromosomes in G_2 not sharing a fragment with $c_{1,i}$. However, one chromosome $c_{2,k}$ in G_2 is the mutated equivalent of $c_{1,i}$, i.e. before all reciprocal translocations, they were the same. Hence we may only sum over all other chromosomes:

$$v_{1,i} = \sum_{\substack{j=1 \\ j \neq k}}^{n} (1 - p_{2,j})^{2tp_{1,i}}$$

.

As we do not know which chromosome in G_2 is this remnant of $c_{1,i}$, Mazowita et al. (2006) propose to average over all chromosomes,

$$v_{1,i} = \frac{n-1}{n} \sum_{j=1}^{n} (1 - p_{2,j})^{2tp_{1,i}} \tag{4}$$

.

The expected number of chromosome pairs with no shared fragments can then be calculated as

$$v_1 = \sum_{i=1}^{m} v_{1,i}$$

.

As there are m chromosomes in G_1 and n chromosomes in G_2, there are $m \times n$ possible chromosome pairings with shared fragments. Therefore, the expected number of chromosomes with shared fragments, the expected dispersion F_{G_1,G_2}, can be calculated as the number of possible chromosomes with shared fragments minus the expected number of chromosomes with no shared fragments:

$$F_{G_1,G_2}(t) = mn - \frac{n-1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - p_{2,j})^{2tp_{1,i}} \tag{5}$$

Furthermore, we define

$$\mathbb{1}_{i,j} = \begin{cases} 1 & \text{if fragment of } c_{1,i} \text{ can be found on } c_{2,j} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

.

We can then calculate the observed dispersion $\widetilde{F_{G_1,G_2}}$ of genome G_1 on G_2 as

$$\widetilde{F}_{G_1,G_2} = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}_{i,j}. \tag{7}$$

Mazowita et al. (2006) propose that the best estimation for t, the reciprocal translocation distance between G_1 and G_2, is $\hat{t}$ which sets equal the observed dispersion and the expected dispersion:

$$\widetilde{F}_{G_1,G_2} = F_{G_1,G_2}(\hat{t})$$

.

By substituting, we get Mazowita et al. (2006), equation 9:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}_{i,j} = mn - \frac{n-1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - p_{2,j})^{2\hat{t}p_{1,i}} \tag{8}$$

.

This can be solved numerically for $\hat{t}$. If we know the number of synteny blocks s and the number of fissions f, we can now calculate the inversion distance $\hat{i}$ by using equation 3.

As discussed in the previous chapter, in our model, chromosomes are selected independently of their size for a translocation. Hence we propose a modified estimator $\hat{t}_{\text{Unif}}$ by setting $\frac{1}{m} = p_{1,1} = \cdots = p_{1,m}$ and $\frac{1}{n} = p_{2,1} = \cdots = p_{2,n}$. Therefore, we can simplify equation 8 to

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{1}_{i,j} = m(n-1)\left(\frac{n-1}{n}\right)^{\frac{2\hat{t}_{\text{Unif}}}{m}} \tag{9}$$

.

In this case, we can solve the equation analytically as

$$\hat{t}_{\text{Unif}} = \frac{m}{2}\frac{\ln\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{1}_{i,j}\right) - \ln(m) - \ln(n-1)}{\ln(n-1) - \ln(n)} \tag{10}$$

.

As before, we can use the modified translocation distance $\hat{t}_{\text{Unif}}$ to calculate the modified inversion distance $\hat{i}_{\text{Unif}}$.

3.2.2 PhylDiag

PhylDiag is a software developed by Lucas et al. (2014) to identify synteny blocks on gene level in a genome-genome comparison. We used it for three reasons: (1) to infer the number of synteny blocks in order to estimate the number of inversion using equation 3, (2) to get a robust measure of the number of chromosome fragments to estimate the number of reciprocal translocations and (3), to get the synteny block size distribution (chapter 6.2).

Distinguishing breakpoints and pseudo-breakpoints

As discussed in the previous chapter, we need the number of synteny blocks to estimate the inversion distance. We can caluclate them based on the number of breakpoints, as shown in equation 2. However, there are a large number of pseudo-breakpoints in our genomes, i.e. breakpoints which are created by gene events instead of chromosomal events. Fig. 4 shows the complicated causality structure in our data.

In combining equation 2 and 3, we get

$$i = \frac{b + \max(c_1, c_2) - \min(c_1, c_2) - f}{2} - t \tag{11}$$

.

If we overestimate the number of breakpoints b by treating all pseudo-breakpoints as breakpoints, we will also overestimate the number of inversions i. Mazowita et al. (2006) faced this problem when they analysed their DNA at high resolutions. In high resolutions, the

noise, i.e. pseudo-breakpoints, increased which lead to overestimated numbers of inversions. PhylDiag filters pseudo-breakpoints by (1) filtering genomes for genes which are only present in one of both genomes, (2) grouping duplicated genes, and (3) allowing gaps in synteny blocks. By filtering out genes which are only present in one genome, it ignores pseudo-breakpoints caused by deletions and gene births. Secondly, by grouping duplicated genes, it filters pseudo-breakpoints caused by tandem duplications. Finally, by allowing gaps in synteny blocks, PhylDiag filters pseudo-breakpoints created by distant duplications. As part of this thesis, we analysed our data to find the optimal parametrisation of PhylDiag to decrease the noise from pseudo-breakpoints. The most important parameter was the allowed gap size in synteny blocks, which was finally set to 1. We found that these artificial breaks of synteny blocks come from distant duplications, again proving that our dichotomous model of tandem duplications and distant duplications is valid.

Chromosomal fragments

PhylDiag was used to retrieve the number of chromosomal fragments of genome G_1 on each chromosome of genome G_2, which was used to solve equation 6.

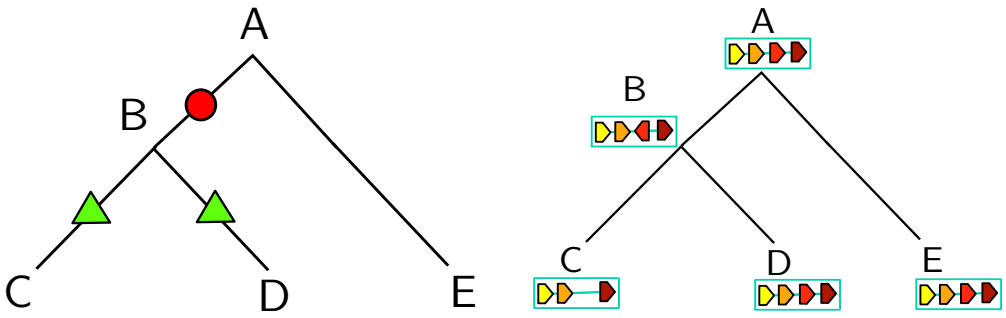
Synteny block size distribution

PhylDiag was also used to measure the sizes of synteny blocks, or with different words, to measure the distances between two breakpoints. The distribution of the sizes was used as a quality measure for our estimate of the number of reciprocal translocations and inversion during our numerical optimization (chapter 6.1). Afterwards, the resulting empirical cumulative density function was used to adapt our inversion size distribution (chapter 6.2).

3.3 From distance to branch length

In the previous chapter, we showed how we measure distances between leafs of the tree (e.g. C-D, C-E and D-E in Fig. 5(a)). However, we need per-branch input for our model and the simulation. Hence, we need a method to estimate per-branch events based on our distance measures. There are several possibilities: statistical approaches, as the least squares estimation, or algorithmic approaches, as neighbour-joining.

In the following, we use the formalism of Felsenstein (2004), pages 151-152. Given n species s_1, ..., s_n and a distance between two species $\| \cdot \|$, we can define the distance vector d as



(a) Problem in placing events

(b) Problem in consecutive mutation cancelling prior mutations out

Figure 5: Exemplary phylogenetic trees indicating the difficulty in branch estimation
Fig. 5(a) shows the problem of non-identifiability in the reconstruction of events. There is no way, knowing only modern species C, D and E, if a specific event occurred once early at position 1 (red circle) or twice at positions 2 (green triangles). Fig. 5(b) shows other error sources, which lead to underestimation of events. An inversion takes place between A and B. Yet, we do not see it in descendant C because of a gene deletion. We do not see it in D either, because another inversion cancelled the first inversion out, thus both inversions cannot be observed.

$$d = \begin{pmatrix} \|s_1, s_2\| \\ \|s_1, s_3\| \\ \vdots \\ \|s_{n-1}, s_n\| \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{\frac{n(n-1)}{2}} \end{pmatrix} \tag{12}$$

.

In our case, the distance $\|\cdot\|$ is the number of reciprocal translocations or inversions between two species. In the example Fig. 5(a), the distance vector would be

$$d = \begin{pmatrix} \|C, D\| \\ \|C, E\| \\ \|D, E\| \end{pmatrix} \tag{13}$$

.

3.3.1 Linear least squares estimation (LM)

To estimate the vector of event numbers per branch v using usual least square estimation, we can follow Felsenstein (2004), equation (11.9),

$$v = (X^T X)^{-1} X^T d \tag{14}$$

.

with X a positive connection matrix (compare Felsenstein (2004), page 151, equation (11.7)). In a connection matrix X, rows represent combinations between leaves in the same order as in d, and columns represent the branches b. In a given row, every branch which lies on the path between the two leafs gets a 1, all others get a 0. Formally:

$$X_{(s_1,s_2),b} = \begin{cases} 1 & \text{if } b \in \text{Path}(s_1, s_2) \\ 0 & \text{else} \end{cases} \tag{15}$$

.

For example, the connection matrix for tree Fig. 5(a) would be

$$X^* = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \tag{16}$$

.

A consequent problem is that the rank of X is not full, and thus the placement of events is non-identifiable. As we cannot observe A, there is no way to decide if an event happened between A and B or between A and E, which leads to equality of column 1 and 4 in X. Our approach was to treat B-E as one branch in the estimation, and afterwards assume an equal rate of events on all of the branch. E.g., if we estimate the distance between B and E to be 5 events, and the phylogenetic distance A-B to be 2 million years, and the phylogenetic distance A-E to be 8 million years, we will place 1 event on branch A-B and and 4 events on A-E.

3.3.2 Weighted linear least squares estimation (LM Weighted)

A popular way to improve the simple linear model is to assign weights to observations, depending on their reliability. In our case, genomes which are closer in the phylogenetic tree are less exposed to noise as gene events, breakpoint reuse, or similar. Therefore, the weight of an observation should be inversely proportional to the phylogenetic distance. Another possible distance would be the number of generations between two genomes, however Huttley et al. (2007) suggest that phylogenetic distance is may be more appropriate.

Given the phylogenetic distance $\|\cdot\|_p$, we define a diagonal weight matrix W as follows:

$$W^{-1} = \begin{pmatrix} \|s_1, s_2\|_p & 0 & \dots & 0 & 0 \\ 0 & \|s_1, s_3\|_p & \dots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & & 0 & \|s_{n-1}, s_n\|_p \end{pmatrix} \tag{17}$$

.

Following Felsenstein (2004), page 151, equation (11.12), we can then estimate the event numbers per branch v as

$$v = (X^T W X)^{-1} X^T W d \tag{18}$$

.

3.3.3 Non-negative least squares estimation (NNLS)

A linear model can result in negative branch estimates (see Sankoff and Mazowita (2005)). Hence, we use non-negative least square optimization. We look for v satisfying

$$\min_v \frac{1}{2} ||Xv - d||_2 \qquad \text{with } x \geq 0 \tag{19}$$

.

Equation 19 is solved numerically with the function NNLS in the framework of the Python library Scipy.Optimize[1]. The results were checked with the results of the R package NNLS (Mullen and van Stokkum (2012)).

3.3.4 Minimum evolution (ME) and Neighbor-joining (NJ)

Other popular approaches to calculate branch lengths from distances are two algorithmic approaches, called *Minimum evolution* and *Neighbor-joining*. They are described in detail in Felsenstein (2004), pp. 159-161, and Felsenstein (2004), pp. 166-171. The Minimum evolution approach creates several phylogenetic tree topologies, estimates the branch lengths on each of those using unweigthed least squares estimation, and finally chooses the tree where the absolute sum of all branch lengths is minimal. Neighbor-joining, on the other side, creates a tree by always joining the two closest species, until all species have joined one tree. Without noise, Neighbor-joining finds the same branch lengths as unweighted least square estimation (Felsenstein (2004), p. 166). Both algorithms create their own phylogenetic tree

[1] `http://docs.scipy.org/doc/scipy/reference/optimize.html`

besides estimating branch lengths. While this is desirable if the phylogenetic structure is unknown, in our case this creates problems. Both Neighbor-joining as well as Minimum evolution do not find the real phylogenetic tree with our inversion distance and reciprocal translocation distance. This is due to the fact that both mutations did not appear evenly distributed over the tree. For example, human and chicken are, according to the reciprocal translocation distance, closer than human and mouse. Therefore, both algorithms create trees where the most common ancestor of human and chicken lived long after the most recent common ancestor of human and mouse. As a result of this wrong tree, we cannot use the resulting branch estimates, as branch lengths for branches above ancestor have no meaning. For completeness, however, we included their fit in our results table 1.

3.3.5 Comparison of methods

	Translocations		**Inversions**	
	R-squared	**LogLikelihood**	**R-squared**	**LogLikelihood**
LM	0.9924	-549.39	0.9916	-1036.81
LM Weighted	0.9931	-519.84	0.9938	-986.56
NNLS	0.9764	-554.27	0.9630	-1041.38
NNLS Unif	0.9784	-525.44	0.9637	-1040.49
NJ	0.9751	-560.16	0.9939	-851.92
Fast-ME	0.9751	-560.16	0.9939	-851.92

Table 1: Estimation method comparison

NNLS: Non-negative least squares estimation. LM: Linear model. LM Weighted: Linear model with weights indirectly proportional to phylogenetic distance. NNLS Unif: Non-negative least squares estimation based on translocation/inversion distances based on a modified Mazowita et al. (2006), equation 9 (chapter 2.4.2). NJ: Neighbor-joining. Fast-ME: Minimum Evolution.

The branch lengths estimated using the methods chapter 3.3.1 to 3.3.3 can be found in annexe, table 7. Least squares estimation and weighted least squares estimation were used on two distances, the translocation distance and inversion distance. Non-negative least squares estimation was used on synteny block distance, translocation distance, the modified translocation distance (NNLS-Unif), inversion distance and the modified inversion distance (NNLS-Unif). For the translocation distance, all 3 methods provide comparable results. However, in some species like Gallus gallus, both least squares estimation and weighted least squares estimation result in negative estimates. In reality, there is no sense in observing -3 inversions on a certain branch. Though it is theoretically possible to reverse chromosomal rearrangements, it is a very unlikely event. We therefore decided to only allow for non-

negative estimates. The same holds true for the inversion distance. However, the differences between the methods are more apparent, as non-negative least squares provides much smaller estimates than least squares estimation in Bovidae (-26 inversions) and Homo sapiens (-8 inversions).

To compare the fit of the models, we calculated R^2 and the log-likelihood. To calculate R^2, we calculated the ratio of the residual sum of squares and the total sum of squares, or as formula:

$$R^2 = 1 - \frac{(d - X\hat{v})^T(d - X\hat{v})}{(d - \overline{d}\mathbf{1}_{\frac{n(n-1)}{2}})^T(d - \overline{d}\mathbf{1}_{\frac{n(n-1)}{2}})}$$

where $\overline{d} = \frac{2}{n(n-1)} \sum_{i=1}^{\frac{n(n-1)}{2}} d_i$, the mean of d, and $\mathbf{1}_{\frac{n(n-1)}{2}} = (1, 1, \ldots, 1)^T$, the vector of ones of dimension $\frac{n(n-1)}{2}$. The log-likelihood was calculated under the assumption of normally distributed residuals.

The results are displayed in table 1. Non-negative least squares has a slightly worse fit than least squares estimation or weighted least squares estimation (lower R^2 and log-likelihood value) for both, translocation and inversion distance. It fits better, however, than Neighbor-joining and Minimum evolution for the translocation distance. Finally, we decided to use non-negative least squares in the rest of our work. Least squares estimation and weighted least squares estimation, as discussed above, resulted in non-interpretable negative coefficients while providing only slightly better over-all fit. Neighbor-joining and Minimum evolution created phylogenetic trees which differed from reality, thus they provided no estimates for inner branches and their estimations could not be used.

Fig. 6 sums up our estimation process (green box) from genome data to reciprocal translocation and inversion numbers per branch.

3.4 MagSimus

MagSimus is a gene order evolution simulation programmed in Python, created by the group of H. Roest Crollius. As input, it takes the modern genomes, the gene content of the ancestral genomes, a phylogenetic tree, and several parameters. MagSimus derives from it a simulated ancestral genome, in our case Amniota. The ancestral genome then evolutes along the branches of the phylogenetic tree. The event rates along the branches and design choices of the operations are specified as parameters.

As part of this thesis, chromosomal rearrangements were implemented following our design decisions in chapter 2.2. However, we also implemented the alternative approaches, e.g.

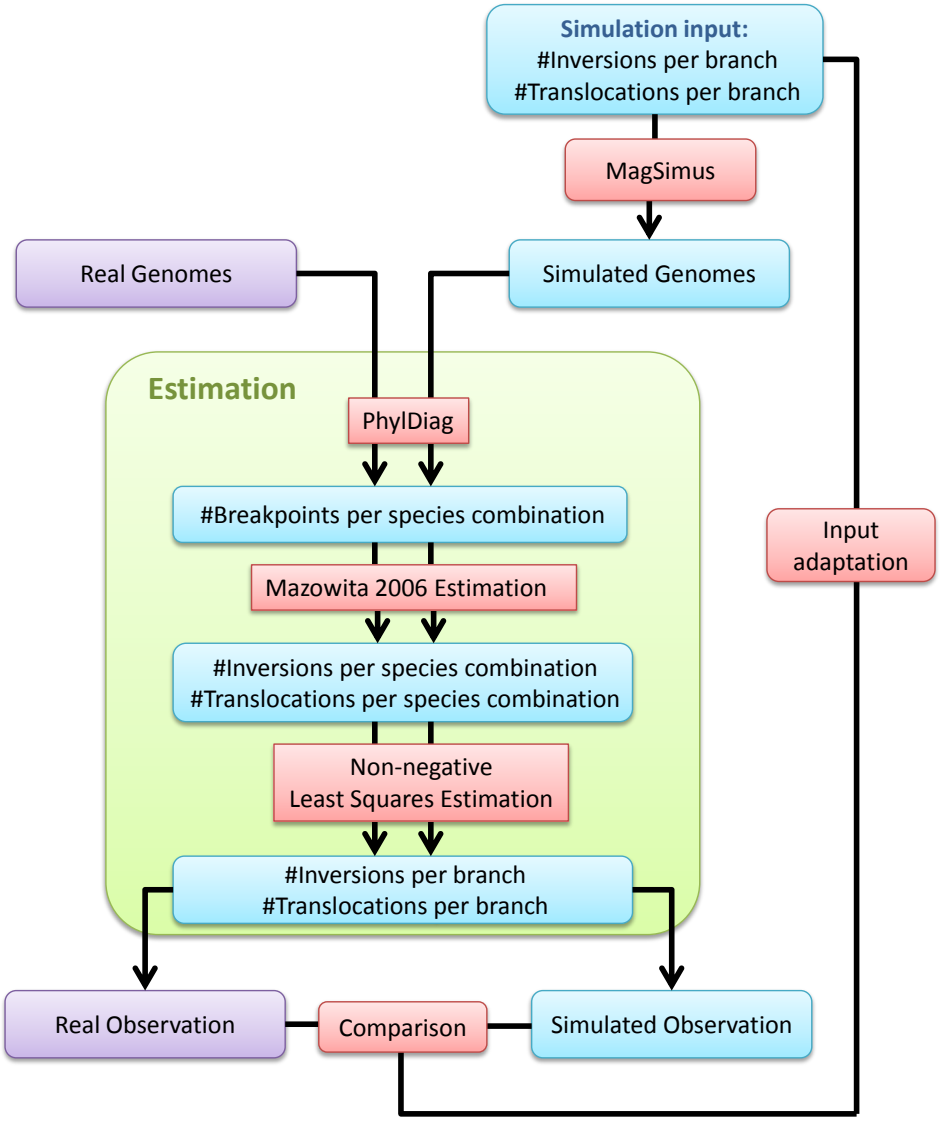


Figure 6: Schematic representation of the pipeline for chromosomal rearrangement estimation In the first step (purple), the real genomes are input in our estimation pipeline. The output, i.e. the real observation in form of the estimated number of chromosomal rearrangements, is then used as initial input parameters for the MagSimus simulation. MagSimus is launched, and its output, simulated genomes, are inserted in the estimation pipeline. The output, the simulated observation in form of the estimated number of chromosomal rearrangements, is compared to the real observation. The input of MagSimus is changed on a branch basis by the difference between real and simulated observations. The simulation process begins again until the simulated observations equal the real ones.

proportional sampling for reciprocal translocations, to test the effects of different models.

As the simulation is rather slow, we decided to use only a subset of our genomes in the simulation. We decided to use 5 species: human, mouse, dog, opossum and chicken (see Fig. 7). They are at different ends of the complete phylogenetic tree and represent opposing

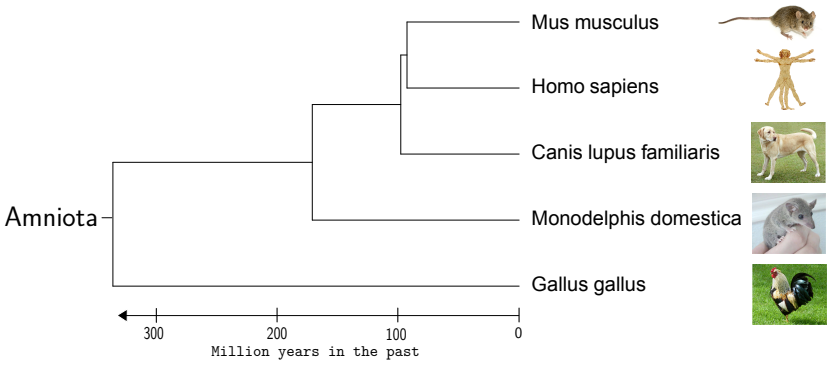


Figure 7: Phylogenetic tree of the species selected for the gene order evolution simulation MagSimus.

genome features. E.g., the opossum has very few, large chromosomes, whereas the chicken as many small micro-chromosomes. Furthermore, for the simulation, we erased the smallest 4 chicken chromosomes, as they consisted of only very few genes and posed subsequent problems in the simulation.

Before we could start the simulation, we needed to specify an additional parameter: the genome at the root of the tree, i.e. the Amniota genome. Starting from this genome, the different mutations can be executed on the branches of the phylogenetic tree to create simulated modern genomes.

Amniota genome: number and sizes of chromosomes

As a result of ChromEvol 2, we set the initial Amniota genome to 21 chromosomes. A direct estimate of the number of genes in the Amniota genome can be gotten from the gene trees, which indicate 20,292 genes for the ancestral Amniota genome. Interestingly, an indirect approach to obtain an estimate of the number of genes is to use a linear regression on the gene number of available Amniota (table 5). Here, we regress the number of genes G in a genome by the number of chromosomes C in the genome. As a result, we get

$$G = 22,390.9 - 144.7C$$

If we use this equation to predict the number of genes in a genome with $C = 21$, as estimated with ChromEvol 2, this results in an estimated 19,352 genes. This is slightly less than the estimate using the gene trees. We decided to use the larger estimate, as there are still genes added to the Ensembl data base, thus our estimation based on linear regression will likely increase in the future. We note, however, that the small difference is not likely to impact the results.

To infer the initial chromosome sizes in Amniota given its number of chromosomes and genes, we averaged over the sizes of modern chromosomes. As most modern species do not

have exactly 21 chromosomes, we used the following interpolation approach. At first, we standardized all modern species genomes both on their number of chromosomes and on their number of gene content on these chromosomes. E.g., the human has 23 chromosomes, thus each one should account for $1/23^{th} = 4.3\%$ of the genes if all genes were distributed equally. In reality, chromosomes have different sizes, as discussed in chapter 2.4.2. For example, the biggest human chromosome accounts for 10.4% of the genes, more than twice as much as in a uniform distribution. Figure 2 shows the cumulative density functions (CDF) for all modern Amniota species. In our model, Amniota had 21 chromosomes, thus one chromosome represents about $1/21^{th} = 4.8\%$ of chromosomes. To get the size of the largest Amniota chromosome, we now averaged all CDFs at 14.8% of the chromosomes, which resulted in about 14.8% of the genes. For the size of the second largest chromosome, we would look at $2/21^{th} = 9.6\%$ of the chromosomes and so forth. The resulting discrete CDF is shown as red dots in the plot.

As can be seen, the first 4 chromosomes of our simulated genome (first 4 red dots from the left) have higher standardised gene sizes than the majority of other genomes (the red dots are above most of the dot cloud for a given x-value). This is due to the fact that 3 species, indicated by the 3 dots over each red dot, seem to have a different CDF shape than the rest of Amniota species, which are closer to the black line. Interestingly, these 3 species are all birds, and this observation is known as the dualism of large macro-chromosomes and small micro-chromosomes in birds. In other words, birds have a few very large chromosomes and many very small chromosomes with just a few dozen genes. This leads to steeper CDFs in our graphic. Our simulated Amniota genome is now a compromise between these extreme bird chromosomes and homogenous mammal chromosomes.

In the final step, we distribute the 20,292 genes by the estimated gene proportion over the 21 chromosomes. E.g., the biggest chromosome should have 14.8% of 20,292 genes, resulting in 3013.14 . As gene numbers per chromosome can only be integers, we used the Hare-Niemeyer method to allocate the genes fairly. In our case, the biggest chromosome of Amniota has 3,013 genes and the smallest 282.

Finally, we simulated genomes for the 5 species, using our inferred event rates as input parameters. However, when we used the same estimation techniques as before to infer the event rates based on the simulated genomes, we did not find the same rates as we used as input to the simulation. We therefore concluded that our estimation pipeline does not provide unbiased estimates for our model (see chapter 5).

4 Data

There are three types of data used for this thesis.

1. Modern genomes for Amniota species, including gene names, gene and chromosome lengths in nucleotides and gene orientations, together with their gene trees, i.e. the inferred history of the genes, taken from the Ensembl 78 data base (Cunningham et al. (2015))

2. Chromosome numbers for genomes stored in Ensembl 78, the Genome Size database (Gregory (2015)) and the GOLD database (Reddy et al. (2015))

3. Phylogenetic trees from Ensembl 78 and `http://timetree.org` (Hedges et al. (2015))

4.1 Genome data

Modern Genomes

There are 50 Amniota species in Ensembl 78 (see Table 5). As discussed before, this thesis focuses on large chromosomal rearrangements. To accurately assess their number and size, it is crucial that the genomes used for the analysis are well assembled, i.e. the gene order indicated by the data must be reliable.

Despite of cheap sequencing technology being available for several years, completely assembled genomes are still rarely available. Instead, most genomes consists of larger and smaller segments, called scaffolds. To place these scaffold on a chromosome in the correct order, is a difficult problem. In our study, we focused our analysis on genomes that contained few such errors.

We want to model gene order, not nucleotide order. Genes are annotated in genome sequences using complex bioinformatic pipelines. This data, including the lengths and positions of genes, where downloaded from Ensembl (Cunningham et al. (2015)). Of all the genes, e.g. 60,000 for humans, only the coding genes are selected, i.e. only genes which are used to code for proteins. This leaves 21,796 in the human genome, excluding the Y chromosome (see Table 5). There are 50 Amniota species datasets available in Ensembl 78. Some of these datasets include scaffolds, i.e. segments whose location on a chromosome is still unknown. In some datasets, like in Monodelphis domestica, several such segments are grouped together to a pseudo-chromosome called Un. To reduce noise in our analysis, we do not consider genes on scaffolds and Un. Furthermore, sex chromosomes Y and W were also not considered. Y is

the small, male sex chromosome in mammals, and W is the small, female sex chromosome in birds. Both chromosomes have in common that they do not have many genes on them and have very altered evolutionary constraints compared to the other chromosomes.

An important question is if chromosome X and Z, the partners of Y and W, should be included, as they also behave quite differently compared to other chromosomes due to their nature as sex chromosomes. As they nevertheless fit well within the size distribution of autosomes and represent a substantial part of the genome, we decided to follow Mazowita et al. (2006) and to include them in our data.

The number of genes eliminated during this selection process is represented in table Table 5. We finally chose to include all genomes where at least 50% of coding genes were conserved in the final genome. A corner case was the green lizard, *Anolis carolinensis*, with about 51% genes eliminated. We first included it as it is the only representative of a large group of animals, and could have greatly helped to accurately describe when and in which lineage certain mutations took place. Finally, however, it was discarded, as descriptive statistics showed that the noise introduced by its inclusion out-weighted the possible information gain.

This selection process left 21 Amniota genomes in our data base. A phylogenetic tree indicating their relationships is shown in appendix, Fig. 15.

Gene trees

Phylogenetic gene trees are formal representations which describe the history of genes. Based on comparison in different species, and based on the number of accumulated mutations such as base changes, the history of a gene can be inferred. This includes when a gene appeared for the first time (gene birth), when it got duplicated (gene duplication) and if it was erased on certain genetic lineages (gene deletion). For example, if a gene can be found in human, chimpanzee, gorilla and orang-utan, but in no other animals, it is most likely that it appeared in a common ancestor of the four primates that they do not share with other animals. Hence, this gene birth event would be dated at the Hominidae lineage (see appendix, Fig. 15).

Due to restrictions in our data, there are two types of events which cannot be observed. Firstly, there may be genes which were present in an ancestor, but no copies remain in modern genomes, thus making them invisible in gene trees. Secondly, in our data, we cannot distinguish certain gene duplications from gene births on edges leading to leaves. For example, let there be a gene birth on the edge between human and the common ancestor of human

and chimpanzee. If this new gene gets duplicated, in our data it will not look like a duplicate but rather as another gene birth, thus artificially increasing the number of gene births and decreasing the number of gene duplications on these outer edges.

These gene trees can be downloaded from Ensembl 78. However, the inferred number of genes in ancestral genomes are likely inflated. Most recent Amniota species have about 20,000 coding genes, yet the original trees indicate up to 35,000 coding genes in one species. Hence, we used edited trees provided by Peres and Roest Crollius (2015).

4.2 Chromosome data

Though gene trees indicate the number and types of genes in extinct ancestral species, we do not know about the evolution of their order or their distribution over chromosomes, which is the reason why the MagSimus simulation was created (see chapter 3.4). One particular information which is missing is the number of chromosomes in an ancestral species. Though there are studies which try to infer the number of ancestral chromosomes as far back as Theria (Deakin and Ezaz (2014)), and Ouangraoua et al. (2011) discusses several chromosomal features of the Amniota chromosome, to our knowledge there are no estimates for the chromosome number of Amniota.

We considered that estimating the number of chromosomes in Amniota based on only the 50 modern Amniota species which are available in Ensembl 78 would be unreliable, as the number of chromosomes in modern species varies considerably over all Amniota. Furthermore, there exists considerable differences in the chromosome size distribution of e.g. birds, which have a few very large macro-chromosomes and many small micro-chromosomes, and Theria species, which have in comparison very even-sized chromosomes. As there are few bird and reptile genomes in Ensembl 78, we considered their sample size too small to draw conclusions on the chromosome number of Amniota. Therefore, we downloaded chromosome count data from the Genome Size database (Gregory (2015)) for all Amniota species represented in this source. Furthermore, we checked, where available, this data with the GOLD database (Reddy et al. (2015)) and Ensembl 78. This resulted in 608 samples. For 10 species, there was contradicting information, which could not be resolved using available resources. As discussed in chapter 3.1, these cases were treated with an equal chance of observation.

4.3 Phylogenetic trees

There are two sources of phylogenetic data used for this thesis. The first comes directly from Ensembl 78. This data is used for the estimation of chromosomal event rates, gene event rates and used in the evolution simulation called MagSimus (see chapter 3.4). Furthermore, to infer the chromosome counts of ancestral species as discussed in chapter 4.2, a phylogenetic tree which includes the Amniota species with available chromosome count was needed. The most complete tree was found at `http://timetree.org` (Hedges et al. (2015)). Timetree is a project that combines estimations on the dating of speciations events from hundreds of studies and tries to build a consistent phylogenetic tree out of it. The 608 Amniota species with chromosome count were matched against the leaves in the phylogenetic tree. 90 of the 608 species were not included in the tree and hence discarded, resulting in a final phylogenetic tree with 518 species.

5 Parameter estimation

In this chapter, we explain how we estimated the number of reciprocal translocations and inversions in our phylogenetic tree. However, as an intermediate step, we need to estimate the number of fusion and fissions.

5.1 Fusions and fissions

We use our phylogenetic tree of 518 Amniota species together with their respective chromosome count to estimate the rate of fusions and fissions. As allosomes cannot be as easily rearranged as autosomes, we used ChromEvol 2 only on the autosome number (chromosome number - 1), and afterwards added one for the sex chromosome.

If we let ChromEvol 2 optimize the fusion and fission rates independently, it estimates them at 0.587 and 0.619, concluding in AIC = 3281.7 . However, this results in an estimate of only 1 autosome in the Amniota genome, indicating a transient state. This is a disappointing and unrealistic result, especially as no modern Amniota genome has so few autosomes. We therefore tried a model which forces both transition rates, i.e. the fusion and fission rate, to be equal. We used a grid search to calculate the AIC for different transition rates. The results are shown in Fig. 8.

The optimal fission and fusion rate was inferred to be 0.525 per million years, with AIC = 3283.5. The AIC difference between the optimal AIC with equal transition probabilities

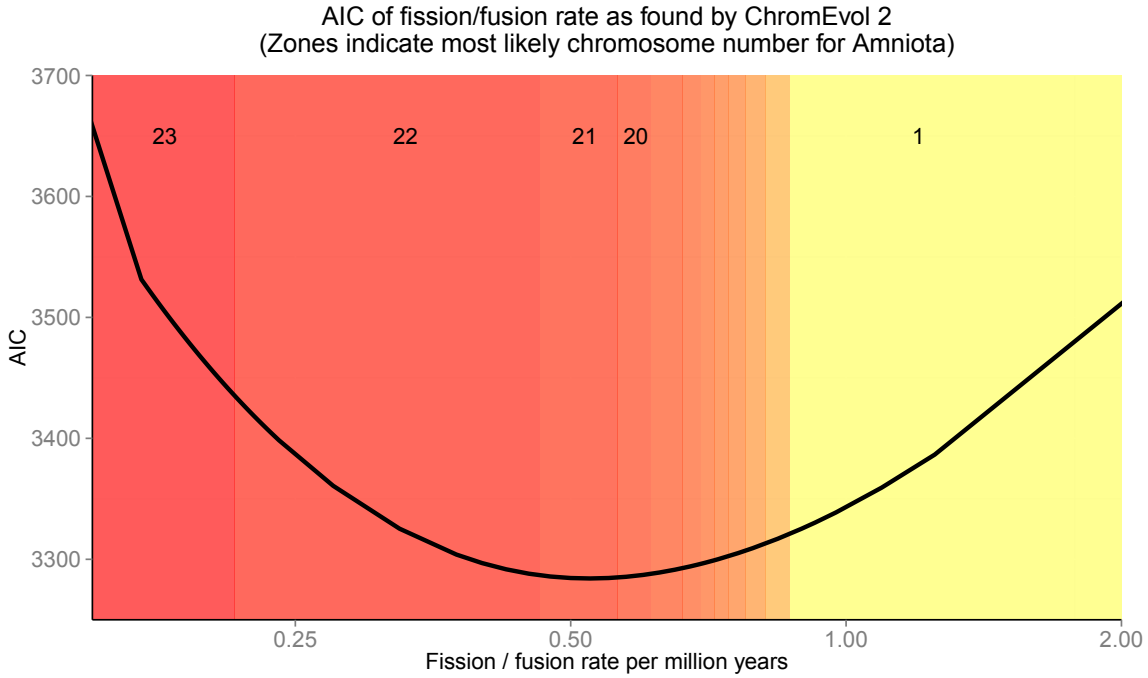


Figure 8: Inferring the optimal fusion/fission rate using ChromEvol 2

AIC for continuous time Markov processes for different fusion/fission rates, calculated with ChromEvol 2. The descending (fusion) and ascending (fission) rates were set to be equal. The minimal AIC was measured at a rate of 0.525 fissions/fusions per million years. The coloured zones indicate the most likely numbers of chromosomes in the ancestral Amniota genome. #AICChromEvol2ForFusFisRatesAmniota

and differing transition probabilities is not big compared to AIC increases due to change in the transition probability (Fig. 8). One factor may be the increase of degrees of freedom by using one parameter less, thus compensating a worse fit. This model leads to an estimate of the number of chromosomes in the Amniota genome of 21, which is seems plausible (see appendix, Fig. 18 for the inferred probabilities π_i for different numbers of chromosomes in Amniota in this model).

However, the fusion and fission rate seems rather high. For example, in the 12 million years between human and chimpanzee, there was only one fusion, thus implying a rate of about 0.08. However, as the rate was constraint to be fix over all the tree, fast mutating genomes like that of rodents increase this estimate. Furthermore, there are lineages where the number of fusions and fissions is not balanced, e.g. the dog branch, where most likely happened many fissions and nearly no fusions. However, as we constrain the model to assume equal fusion and fission rates, the model will consider more fusions than happened in reality, thus further increasing the estimate.

The results of Deakin and Ezaz (2014) imply that fusions and fissions are rare events,

and using our estimated fusion and fission rate to calculate branch event numbers ill result in unrealistic large estimates compared to cytogenetic results. Besides that, the estimated event numbers on the branches returned by ChromEvol 2 were wrong[2], thus reducing the confidence in the ChromEvol 2 results.

However, we found the estimates of the ancestral chromosome numbers to be realistic and in line with previous estimates. For example, Deakin and Ezaz (2014) find 19 chromosomes in the Theria ancestor, where our model estimates 18. The other estimates were either exact or also in the range of ± 1 chromosome. In order to get event numbers on each branch of the tree based on these realistic values, we constrained the model to only execute either fusions or fissions on each branch of the tree. For example, if the ancestral species A has 19 chromosomes, and the modern species has 21, than the model will execute 2 fissions. Secondly, we reduced the phylogenetic tree with 518 species to our tree of 21 (or 5) Amniota species. For example, there are three intermediate branches between Mus musculus and Euarchontoglires in our tree of 21 species (appendix, Fig. 15), which are not present in our 5 species tree. For the reduced tree, we cumulated the estimations of these 3 branches. This procedure should give very rough estimates. However, if comparing to previous estimates, our estimates prove to be realistic (compare to estimates of Zhao and Bourque (2009)). The results of this procedure are shown in table 2.

	MagSimus branches	**Rate per mya**	**Number of events**
Fissions	Boreoeutheria	0.0845	6
	Canis lupus familiaris	0.1684	16
	Euarchontoglires	0.2000	1
	Gallus gallus	0.0399	13
	Homo sapiens	0.0222	2
	Monodelphis domestica	0.0422	7
	Mus musculus	0.0111	1
	Theria	0.0000	0
Fusions	Boreoeutheria	0.0000	0
	Canis lupus familiaris	0.0105	1
	Euarchontoglires	0.0000	0
	Gallus gallus	0.0153	5
	Homo sapiens	0.0444	4
	Monodelphis domestica	0.0964	16
	Mus musculus	0.0667	6
	Theria	0.0188	3

Table 2: Branch estimates for fissions and fusions for a phylogenetic tree of 5 Amniota species

Finally, we estimated the same parameters, but based on a fusion and fission rate of 0.1

[2]A bug report was filed.

per million years which is closer to the rate found in the human lineage, for example. This lead to an estimate of 23 Amniota chromosomes and an AIC = 3842.6, which is much larger than from our previous results. Even more, the chromosome number of ancestral species where previous estimates existed were found to be too large, hence we stayed with our previous estimates.

5.2 Reciprocal translocations and inversions

As discussed in chapter 3, we first calculate the reciprocal translocation distances between all genomes, using PhylDiag and the estimation method of Mazowita et al. (2006). Using this result and another output of PhylDiag, we can calculate the inversion distances between all genomes. Finally, we use non-negative least squares to estimate the the lengths of the branches in reciprocal translocations and inversions. Our results can be found in table 7. Notably, both the original estimator proposed by Mazowita et al. (2006) and our modifed version (equation 9) find nearly identic results. The modified estimates are slightly better to fit on the tree using NNLS (table 1). However, this difference is so small that even slight changes in the genes considered for our analysis allowed the original estimates to fit better. As the differences are so small (also in the later simulations), we present further results only based on the original estimator which represents a model more accepted in the literature.

However, our only measures of the accuracy of these results were the previously discussed R^2 and Log-Likelihood, which only analysed the fit of the branches given the distances. However, this does not take into account possible errors in the estimation of the distances. In order to evaluate the estimator proposed by Mazowita et al. (2006) and to measure the quality of PhylDiag, we decided to use the MagSimus simulation.

As discussed in chapter 3.4, MagSimus is rather slow, and we therefore decided to reduce the number of species in the simulation to 5. To check in how far this reduction changes our estimation, we compared the estimates only based on the 5 genomes (table 3, row NNLS-MS) and the estimates based on all 21 genomes, but reduced to the small tree by accumulating intermediate estimates as described in chapter 5.2 (table 3, row NNLS-Shrink). All differences in translocations estimates are small. As a comparison, one can consider the numbers in parentheses in line *OI + NNLS + M2006 + PhylDiag (CI)*. They indicate the simulated 95% interval for a simulated number of translocations given in row *Optimal input*. Both, our estimates based on the 5 genome tree as well as on the 21 genomes tree are well within these margins. Hence we conclude that we do not overfit the number of translocations in our

small tree more strongly than in the tree with 21 genomes. For the number of inversions, however, the estimates based on the 21 genomes tree are outside the interval for Boreotheria, Euarchontoglires and Gallus gallus. This may indicate an overfitting in the estimates based on the 5 genome tree, and estimates from the numerical analysis for these branches should be analysed cautiously.

		Boreo-eutheria	**Canis lupus**	**Euarchon-toglires**	**Gallus gallus**	**Homo sapiens**	**Monodelphis domestica**	**Mus musculus**	**Theria**
Translocations	NNLS-Shrink	19.18	25.05	0.00	3.07	4.93	8.37	53.03	2.23
	NNLS-MS	18.23	26.18	0.07	3.17	2.72	6.79	55.35	1.56
	Start input	18	26	0	3	3	7	55	2
	Optimal Input	15	29	0	3	3	3	48	1
	OI + NNLS	15.00	29.00	0.00	2.68	3.00	3.00	48.00	1.32
	OI + NNLS + M2006 (CI)	19.38 (12.96,26.23)	28.32 (21.21,34.92)	1.94 (0,6.32)	3.26 (0,7.31)	2.25 (0,7.02)	7.35 (0.35,17.73)	56.63 (47.49,70.51)	1.60 (0,3.59)
	OI + NNLS + M2006 + PhylDiag (CI)	18.16 (11.34,24.30)	26.31 (20.05,33.22)	1.98 (0,6.21)	2.95 (0,6.80)	2.60 (0,7.39)	7.31 (1.05,15.97)	54.32 (44.50,65.60)	1.45 (0,3.34)
Inversions	NNLS-Shrink	50.64	92.06	8.91	188.84	94.16	213.90	92.20	65.62
	NNLS-MS	75.83	86.59	0.00	176.32	87.03	213.63	89.57	86.54
	Start input	76	87	0	176	87	214	90	87
	Optimal Input	97	86	0	304	98	201	130	150
	OI + NNLS	97.00	86.00	0.00	304.53	98.00	201.00	130.00	149.47
	OI + NNLS + M2006 (CI)	72.99 (62.9,82.71)	82.36 (76.26,89.89)	0.12 (0,1.74)	254.32 (246.62,264.29)	98.31 (91.38,103.79)	189.55 (178.96,199.35)	108.74 (94.74,120.48)	124.82 (121.4,129.71)
	OI + NNLS + M2006 + PhylDiag (CI)	74.95 (59.87,88.13)	85.85 (77.14,95.99)	0.1 (0,2.1)	176 (165.42,187.36)	87.75 (80.35,97.5)	212.72 (196.85,224.35)	89.62 (76.89,100.86)	86.38 (81.19,91.96)

Table 3: Branch estimates for reciprocal translocations and inversions for a phylogenetic tree of 5 Amniota species

NNLS-Shrink: NNLS on all Amniota species, tree shrank to include only MagSimus species. NNLS-MS: NNLS on MagSimus species, using genomes from MagSimus dataset (micro-chromosomes excluded). Start input: rounded NNLS-MS; also MagSimus parameters before numerical optimization. Optimal Input: MagSimus input after numerical optimization; also our numerical estimates. OI + NNLS: NNLS on MagSimus input. OI + NNLS + M2006: Using Mazowita 2006 to get translocation and inversion estimates based on simulated synteny blocks, simulated 95% intervals in parentheses. OI + NNLS + M2006 + PhylDiag: Using Mazowita 2006 to get translocation and inversion estimates based on observed synteny blocks using PhylDiag, simulated 95% intervals in parentheses. See chapter 6.1 for a detailed discussion.

As MagSimus is an implementation of our model, simulated genomes using our estimated mutation rates should resemble the real genomes. To compare the simulated and real genomes, we used the same estimation methods to infer the number of reciprocal translocations and inversion on the simulated genomes (Fig. 6).

However, our estimations based on the simulated genomes were quite different to our previous estimates, thus questioning our previous results. For example, we estimated the

inversion distance between mouse and chicken to be 429 based on our real genomes. However, when using these rates to simulate genomes, we only estimated 309 inversions based on the simulated data.

This underestimation is a result of gene deletions. Most inversions are small, often including only one gene. If a gene gets inverted and later deleted, we cannot observe this inversion, thus leading to underestimation of the true number of inversions. This influence of gene events on our estimation is neither accounted for in equation 8 nor equation 3.

In a first approach, we tried to correct our estimates. Let $\hat{i}_r$ be the number of inversions estimated on the real data, and $\hat{i}_s$ be the number of inversions estimated on the simulated data. The proportion of inversions lost due to gene events is therefore $r = \frac{\hat{i}_s}{\hat{i}_r}$. Hence, we should simulate $i = \frac{\hat{i}_r}{r}$ to observe $\hat{i}_r$ in the simulated data.

However, when this corrected estimate was used as input to our simulation, we still did not observe the correct inversion and translocation distances. Even more, sometimes $\hat{i}_s > \hat{i}_r$, i.e. we overestimate the true inversion distance. This indicates that we were not always able to completely filter all pseudo-breakpoint, discussed in chapter chapter 3.2.2. Therefore, in the third major part of this work, we created a numerical optimization framework in Python to find the correct estimate of inversions (translocations) according to the simulation. The process is described in the next chapter.

6 Optimization framework

In the last part of this thesis, we programmed an optimization framework for the gene order simulation MagSimus.

At first, we created a score object which reduces the complex genome data into analysable statistics. This includes 9 different distances for every genome-genome comparison and 15 branch statistics for each branch. For our simulated data, we calculated 26 additional statistics to keep track how much effect each input has on different observations.

Secondly, we created a score comparison object that compares two scores and calculates possible differences. This was needed in order to decide if recent changes made our simulated data more realistic or not. We also created an object which is able to calculate summary statistics and graphics for all statistics.

Thirdly, we implemented an optimization object which launches MagSimus multiple times with different inputs and calculates the scores. It has two different modes: (1) it calculates scores for a grid or a random sub-sample of a grid for multiple MagSimus input parameters,

e.g. number of inversions or translocation chromosome sampling method. (2) it calculates scores for input parameters drawn from a distribution. The latter is particularly interesting to analyse input parameters with Approximate Bayesian Computation. Furthermore, J. Lucas provided an application programming interface for Python to access large computer clusters via the job scheduling software HTCondor[3]. By accessing this interface, the optimization object can (1) launch MagSimus several times on a cluster, (2) retrieve the results, (3) calculate the mean statistics, (4) adapt the input to MagSimus, and restart from step 1.

We used the latter to calculate the optimal input number of reciprocal translocations and inversions (chapter 6.1). Furthermore, we calculated the scores for 200 different inversion distributions (chapter 6.2). The optimal inversion distribution was used to calculate the estimates shown in table 3.

Finally, we propose an entropy score to both check the quality of our simulated genomes, and possibly measure over-fitting due to the optimization (chapter 6.3).

6.1 Reciprocal translocation and inversion number

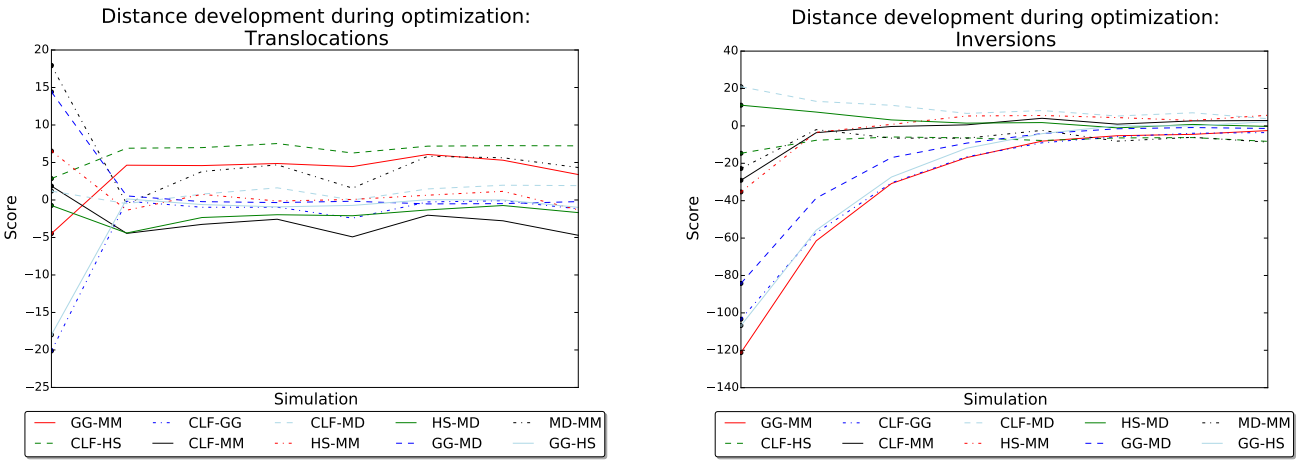
We used the optimization framework to calculate the optimal number of reciprocal translocations (inversions) such that our simulated genomes resemble closely the real genomes. Particularly, we compared two statistics between both datasets: the number of reciprocal translocations and inversions estimated per branch and the distance between the synteny block distributions.

6.1.1 Calculating the numerical estimates

At first, we used our statistical estimates on real data (table 3, NNLS-MS) as input parameters for MagSimus (table 3, Start Input). We then re-estimated the number of inversions and reciprocal translocations per branch based on the simulated genomes. We then added the difference in estimates to our original inputs. We continued this process until we had converged, i.e. the branch estimates on our simulated data stayed the same. Fig. 3.2 visualizes this process.

Fig. 9 shows the convergence process in reciprocal translocation and inversion distances. Both distances converge well to the real observed values. That they do not reach zero for all comparisons may be for two reasons:

[3] `https://research.cs.wisc.edu/htcondor/`



(a) (b)

Figure 9: Convergence process of estimated mean translocation and inversion distances
Each line indicates the difference between real and simulated genome-genome distance. Due to adaptation of the input parameters, the simulated genomes get better to the right, and hence the genome-genome distance gets closer to the value observed in real genomes, thus the difference converges against 0. For every input factor set, 100 simulations were executed, the solid lines indicating the mean over all replications.

1. The variance of the observed values is high, as indicated by the observed extreme values (dotted lines), thus 100 replicates may not be enough to let the mean converge smoothly. We cut the convergence process when the difference stopped decreasing. This was the case after 7 steps at maximum.

2. There may be errors in the measurements of the real distances. We have more observations (genome-genome distances) than fitting parameters (branch lengths), hence if the distances observed in the real genomes are contradictory, our model can only fit up to a certain degree.

The input values of the simulation after convergence can be found in table 3, Optimal Input (OI). For reciprocal translocations, the start input and optimal input are very close, indicating that the estimator of Mazowita et al. (2006) performs well in our model with noise. However, the number of inversions seem heavily underestimated by our statistical estimation. The only exception is Monodelphis domestica, where we overestimated the number of inversions.

To understand how much of the error was contributed by which part of our estimation process, we divided the estimation process in the several individual steps.

6.1.2 Analysing the sources of miss-estimation

Non-negative least squares estimation (NNLS)

table 3, OI + NNLS shows the estimates of the NNLS estimator directly based on our input values. As can be seen, the NNLS estimator does not introduce much under- or over-estimation. Remarkably, the statistical estimates for Theria and Gallus gallus are close to our numerical estimates, even though we forced the NNLS to assume equal rates on both branches. As there is no chance involved, no confidence intervals can be calculated.

Estimation based on Mazowita et al. (2006)

In the next step, table 3, OI + NNLS + M2006, we analyse the simulated chromosomal dispersion (translocations) and the simulated breakpoints. The indicated values are the mean of 100 replications with same input values, and the numbers in parentheses indicate the 95% interval observed. In real genomes, we can only empirically measure both using PhylDiag. However, in a simulation we know exactly what happened, and thus we can easily observe the correct values in the simulated chromosomes. We can see that this doubles the estimated number of reciprocal translocations for Monodelphis domestica to 7.3. However, the confidence interval (CI) ranges from approximately 0 to 17, thus marginalizing this difference. The reason for this is the large number of fusions in the Monodelphis domestica branch. If two chromosomes are included in a translocation and get fused afterwards, we will not see the translocation in our modern genomes, indicating why sometimes 0 translocations are estimated. In the opposite case, one of the two chromosomes implicated in a translocation gets fused with several other chromosomes, thus increasing the relative chromosomal dispersion in the genome, and we overestimate the number of translocations. Overall, there are no relevant miss-estimations, thus confirming our former conclusion that the estimator of Mazowita et al. (2006) for reciprocal translocations performs well. However, our results show that we significantly underestimate the number of inversions in 5 branches (Boreoeutheria, Gallus gallus, Monodelphis domestica, Mus musculus, Theria). As this error was not present in the estimates based on the NNLS estimator alone, and not in the estimates of the number of translocations, it shows that equation 3 does not work well to infer the number of inversions. The information loss due to gene deletion creates this important underestimation. However, a straight forward correction for it proves to be difficult, as discussed in the previous chapter. This can be an interesting future research topic.

PhylDiag

Finally, instead of retrieving the synteny blocks directly from MagSimus, we observe them empirically using PhylDiag (table 3, OI + NNLS + M2006 + PhylDiag). The use of PhylDiag does not change our estimates considerably in most cases. However, in the inversion estimates of Gallus gallus, Homo sapiens, Mus musculus and Theria, the use of PhylDiag leads to a considerable underestimation of inversions. This is most likely due to excessive filtering of pseudo-breakpoints, i.e. too many breakpoints are filtered out, thus decreasing the number of inferred inversions. On the other side, PhylDiag leads to an overestimation of the number of inversions in the Monodelphis domestica branch. This can be explained by the large number of distant duplications in this branch (table 6). Distant duplications cannot be easily identified as pseudo-breakpoints, and hence they increase the number of breakpoints found by PhylDiag. This again leads to an overestimation of inversions as observed. All of these miss-estimations are outside the confidence interval of 95% of the estimates, thus indicating a systematic error. In future work, direct estimates of the confidence intervals of our numerical estimations (table 3, Optimal Input) may be calculated based on the uncertainty in the observed values.

Furthermore, these last estimates are used for our numerical optimization. Therefore, these estimates are close to the estimates based on the real data (table 3, NNLS-MS). Indeed, they are nearly identical, thus underlining the quality of the simulation and our optimization framework. This proves that the framework can correct for both, under- and overestimation caused by our estimation pipeline. Thus, table 3, Optimal input, can be considered robust numerical estimations of the number of chromosomal rearrangements in these 5 Amniota species.

The 95% confidence interval indicates that there is much variance in the simulation of these genomes. It may be possible to calculate the implied confidence intervals for our numerical estimate table 3, OI, by analysing the variance in the simulation output. This can be an interesting project for future research.

6.2 Inversion size distribution

The last part of our model in need of optimization is the inversion size distribution. After the numerical optimization, we wanted to check the quality of our simulated genomes by measuring another, more independent statistic. Following the literature, we chose the synteny block size distribution.

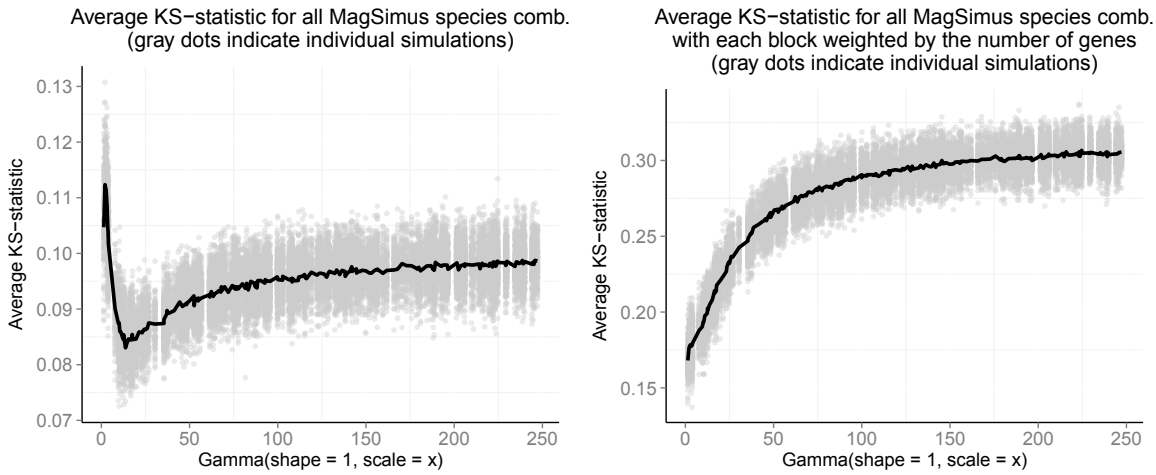
Using PhylDiag, we can measure the sizes of synteny blocks. Based on this observation, we inferred the synteny block size distribution. To evaluate the quality of our simulated genomes, we measured the distance between the synteny block size distribution in real genome-genome comparisons and in simulated genomes. The Kolmogorov-Smirnoff two-sample test is a possibility to quantify this distance. The underlying statistic looks for the supremum of differences in the sample cumulative density functions. However, there are two possible ways to calculate the cumulative density function. The first calculates the density relative to the number of synteny blocks. For example, 1 synteny block accounts for 0.2% in a 500 synteny block comparison. We call the resulting Kolmogorov-Smirnoff statistic the block based KS statistic. However, one synteny block may be large and account for 400 of 20,000 genes in the genome. Thus, it may be reasonable to set it to 2% of the distribution. We call this the gene based KS statistic (see Fig. 22 for a graphical comparison).

The difference between the two is the different weighting of the distributions tail. In genome-genome comparisons there are a few very large synteny blocks. In the block based KS statistic, these represent only a small fraction of the overall distribution, thus the difference between both distributions cannot be large. However, in the gene based KS statistic, these blocks represent a notable proportion of genes, hence the maximum difference between the real and simulated distribution may be achieved in larger synteny block sizes.

Due to our optimization of the number of reciprocal translocations and inversions, we indirectly also optimized the number of synteny blocks. Therefore, we arrive nearly at the same number of synteny blocks as observed in real genomes. However, in real genomes the size of the synteny blocks is often larger than in our simulated genomes. Thus, we should adapt our model to increase the size of synteny blocks. One possible parameter to influence this variable is the synteny block size distribution.

However, it is not obvious how the inversion size distribution should be modelled. Many small inversions create many small inverted synteny blocks. However, between these small synteny blocks are long regions without any breakpoints. Thus, many small inversions will create a synteny block size distribution with many very small and some very large blocks. On the contrary, an inversion distribution with medium sized inversions will lead to a more homogeneous synteny block size distribution.

In order to optimize the fit of the inversion size distribution to our data, we used again our optimization framework. At first, we chose to model the inversion size distribution in accordance to the literature as a Γ-distribution. Due to time constraints, we fixed the shape



(a) KS statistic based on synteny blocks.

(b) KS statistic based on genes.

Figure 10: Influence of the inversion size distribution on synteny block size distribution fitting between real and simulated genomes

The fit of the synteny block size distribution between real and simulated genomes is measured as average Kolmogorov-Smirnoff statistic over 10 genome-genome comparisons (dependent variable). Different inversion size distributions of the form $\Gamma(1, x)$ are tested to increase the fit. Gray dots indicate single simulations, the black line shows the average score of all replications. Fig. 10(a) calculates the Kolmogorov-Smirnoff statistic based on synteny blocks, Fig. 10(b) calculates it based on genes.

parameter of the distribution to 1, to reduce the space on which we had to search. Therefore, our inversion size distribution can also be described as an exponential distribution. This allows still for flexible modelling and is in accordance with previous studies, e.g. Pevzner and Tesler (2003a). We sampled 200 different scale parameters from a uniform distribution between 2 and 250. For each sampled parameter, we used our optimization to converge to the optimal number of inversions and reciprocal translocations in 7 steps, each with 100 replications. Each optimization for one parameter takes about 50 minutes on the cluster of the bioinformatic institut of the ENS Paris. For the 100 replications with the optimal parameters, we calculate the mean Kolmogorov-Smirnoff statistic for all 10 genome-genome comparisons in our 5 species tree. The results are shown in Fig. 10. The statistical properties this averaged value are difficult to assess. However, we found that it represents well the observations in individual genome-genome comparisons. If the scale values get ordered by the size of the associated mean Kolmogorov-Smirnoff statistic, we found that the ranking of the scales were the same in the averaged Kolmogorov-Smirnoff statistic and all individual statistics. Therefore, the average Kolmogorov-Smirnoff statistic seemed to be a good choice for dimension reduction.

The two different methods to calculate the Kolmogorov-Smirnoff statistic lead to two different estimates for the optimal inversion size distribution. The synteny block based Kolmogorov-Smirnoff score arrives at a global minimum at 13.69, whereas the gene based Kolmogorov-Smirnoff score has a boundary optimimum in 2. Due to observations in real data, we know that there are inversions of more than 100 genes. With a $\Gamma(1, 2)$, these are even more unlikely than with a $\Gamma(1, 13.69)$. Therefore, we choose the latter as optimal inversion size distribution (see Fig. 23). Furthermore, we analysed the output with approximate Bayesian computation, using the R package *abc* (). A rejection algorithm with a 10% acceptance rate (tolerance). This resulted in slightly higher scale values (mean = 21.4, median = 28.6). However, the underlying data was considered not sufficient to rely on this estimate (95% confidence interval [9.6, 87.3]).

As discussed in chapter 2.4.3, inversions are often even smaller than a gene. Both distributions set much weight on these small inversions and therefore confirm the previous knowledge. However, the tail of our distributions may be too thin, as in reality we can observe large inversions, however they are very unlikely in our distributions. There are two possible ways to improve upon our results: at first, one can consider other parametric distributions than $\Gamma(1, x)$, which incorporate both, a heivy weight on small inversions and a fat tail. Secondly, the Kolmogorov-Smirnoff statistic may not be optimal to fit the inversion size distribution. As it looks at the maximum distance between the distributions, and most synteny blocks are small, the position of the maximum distance is always lower than synteny blocks of size 10. Thus, the tail, i.e. synteny blocks with sizes above 30 genes, is not fitted at all. We tried to cope with this problems by defining the Kolmogorov-Smirnoff based on genes. However, the effect was not strong enough. Therefore, a different statistic which minimizes the integral between both distributions may be applied more successfully.

6.3 Entropy score

To verify our results and to prevent overfitting, we introduced another score which measures the distribution of genes over chromosomes. The overall problem lies in reducing dimensions of the data. We can look at a dot plot of our real and simulated data and see problems, yet in order to automatically optimize our model, we need a low dimensional score to evaluate our simulation.

We wanted to measure at the same time if the number of distant duplications, the number of translocations and the size of translocations were appropriate. One direct way to summa-

rize them all is to measure the entropy in the data. A genome-genome comparison can be expressed as a contingency table, where each cell is a chromosome-chromosome comparison, and the number within is the number of shared genes. In order to create an entropy measure based on this table, we undertook the following steps.

At first, we cleaned both genomes of genes which were only present in one of both. Secondly, we want to compare the number of observed genes in a cell with the number of expected genes in this cell. This can be done with a χ^2 statistic, and hence can be calculated as follows. Let n_k be the number of chromosomes in genome k, $M_{i,j}$ is the observed number of genes in the comparison of chromosome i of species 1 and chromosome j of species 2, $M_{i,\cdot}$ is the number of genes in chromosome i of species 1, $M_{\cdot,j}$ is the number of genes in chromosome j of species 2, and $M_{\cdot,\cdot} = M$ is the total number of genes in either genome. Hence, the expected number of genes can be written as

$$\frac{M_{i,\cdot} \times M_{\cdot,j}}{M}$$

Therefore, we can define the χ^1 statistics as

$$\chi^2 = \sum_{i}^{n_1} \sum_{j}^{n_2} \frac{\left(M_{i,j} - \frac{M_{i,\cdot} \times M_{\cdot,j}}{M}\right)^2}{M}$$

χ^2 follows approximately a $\chi^2(d)$-distribution, where $d = (n_1 - 1) \times (n_2 - 1)$ are the degrees of freedom.

In order to calculate theses statistics, however, it is necessary that both genomes have the same number of genes, i.e. $M_1 = M_2$. In other words, the sum of the marginal totals must be the same for columns and rows. In our data set this is rarely the case, however, as there are many duplicated genes. E.g., the human genome has one copy of a gene A in chromosome 1, but the mouse has two copies of this gene, one in its chromosome 1, and one in its chromosome 3. Thus, the total number of genes in human is 2, whereas in mouse it is 1.

In order to circumvent this problem, there are 3 possible approaches. The easiest is to delete all genes which are duplicated in one of both genomes. A second approach would count all duplicates as a normal gene, and increase the weight of the unduplicated gene accordingly. In our example, this would mean that both duplicates in the mouth genome count for 1, and the one copy in the human genome counts 2. A third option would be to down-weight duplicates proportionally to their number. In our example, this would mean

that both mouse genes count as 0.5, and the human gene counts as 1.

We compared all 3 approaches. In our data, ignoring all duplicates (approach 1) or down-weighting them (approach 3) resulted in nearly the same p-values, whereas counting them all normally (approach 2) decreased the p-value. However, neither had an impact on the order of p-values, i.e. two very distant genomes always produced a bigger p-value than two closer genomes. Therefore, we choose to take the computationally easiest approach and dropped all duplicated genes (approach 1).

Instead of the χ^2-statistic, we also considered the use of the statistic of the G-Test. Hoey (2012) discusses the derivation of the G - Test and shows that the χ^2 test can be viewed as a Taylor series approximation of the G - Test. Using the formalism from above, we can define

$$G = 2\sum_{i}^{n_1}\sum_{j}^{n_2} \frac{M_{i,j}}{M} \log\left(M_{i,j}/\frac{M_{i,\cdot} \times M_{\cdot,j}}{M}\right)$$

.

G follows the same χ^2-distribution as the χ^2 statistic above (see Hoey (2012)). One drawback in the use of G is the possibility that $\log\left(M_{i,j}/\frac{M_{i,\cdot}\times M_{\cdot,j}}{M}\right)$ can become not well defined for the case of $M_{i,j} = 0$. Of course, for $\frac{M_{i,\cdot}\times M_{\cdot,j}}{M} = c \in \mathbb{R}^+_{\setminus\{0\}}$,

$$\lim_{x\to 0} x \log\left(\frac{x}{c}\right) = 0$$

.

However, as we want to use the statistic as a score, it is unsatisfactory to have the same value for different c. We would like to discriminate between a cell were we expected 100 genes but saw none, and a cell were we only expected 1 gene and then saw none, the latter seeming far more likely given that we only have a finite number observations M and only integer values per cell. Therefore, we added a pseudo-count of 1 to all cells, i.e. instead of using directly $M_{i,j}$ for every cell, we use $M_{i,j} + 1$. This strategy is often used in Bayesian frameworks if such corner cases pose problems. The resulting p-value will not be correct, but can nevertheless work as a mean to compare. We compared the real p-values and the pseudo-p-values after introducing the pseudo-count. As expected, the pseudo-p-value is, in our data, always greater than the real p-value. Intuitively, this can be explained as follows: the G-statistic measures the information content in the table, which can be seen by comparing its formula to the Shannon entropy $-\sum_i p_i \log(p_i)$. By increasing every cell by 1, we are increasing the relative equality of all cells, thus leading to increased entropy. This is equivalent to reducing information content, thus leading to bigger p-values. We checked the

effects of the pseudo-count in all 3 different G-statistic duplications approaches. In no case did the addition of the pseudo-count change the order of the statistics, hence we use it from here on.

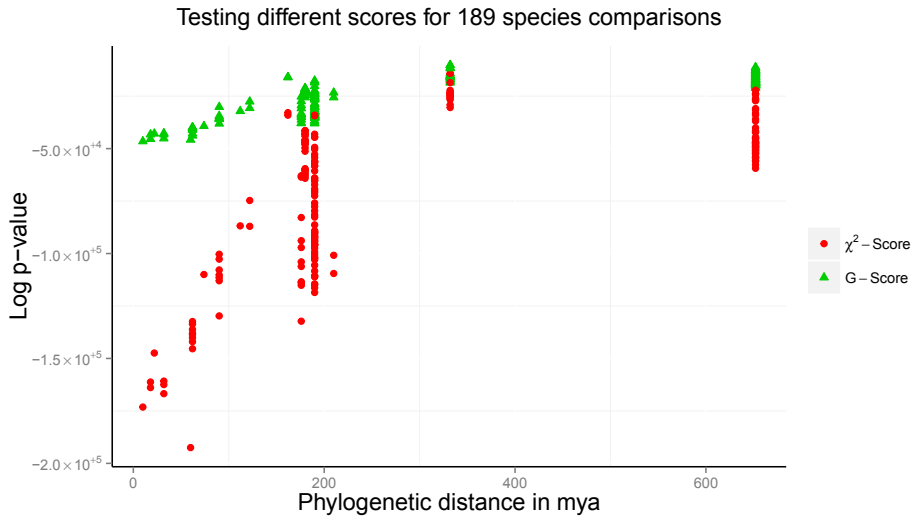


Figure 11: Comparison of two different dispersion measures

In order to calculate the score, all duplicated genes were erased from the genomes. Furthermore, a pseudo-count of 1 was added to every cell in the contingency table, in order to prevent having cells with 0 counts. The G-score seems to be a more reliable due to its smaller variance and it monotonously rising with the distance between two genomes. #EntropyScoreComparisonsForGenomeAnalysis

Fig. 11 shows the results for p-values created with the no-duplication χ^2- and G-log-p-values. Both, the greater variance as well as the inconsistency in the χ^2-score falling with increasing phylogenetic distance lead to the decision to use the G-score to measure entropy in genome comparisons.

As the number of chromosomes in all species were fixed over the course of our simulations, we did not need to translate the G-statistic into a pseudo-p-value. The latter can be used if there is a need to compare tables of different sizes, thus leading to different degrees of freedom for the χ^2-distribution. However, we only compared same sized tables in our case, i.e. the observed real human-mouse gene table with the observed simulated human-mouse gene table. As both the number of genes and the number of chromosomes were fixed in our simulation, only the repartition of the genes in the table changed, thus making it possible to directly compare the statistics.

In our case, we used the ratio of both statistics, which follows approximately a F-distribution. A F-distribution is defined as

$$F = \frac{X_1/d_1}{X_1/d_1}, \quad X_1 \sim \chi^2(d_1), \quad X_2 \sim \chi^2(d_2)$$

In our case, $d_1 = d_2$, hence the ratio follows a $F(d_1, d_1)$ distribution. In 100 replications for all 10 species combinations, this ratio always was in the interval [0.22, 0.79], with a mean of 0.57 in a negatively skewed distribution. To obtain a meta-score of all 10 MagSimus species combination, we used the geometric mean of all 10 ratios. This can be used as a first approximation, as in our case the degrees of freedom are similar in different genome comparisons. An alternative, more valid approach would be to transform the ratios in p-values using the F-distribution, and creating a meta-score based on those.

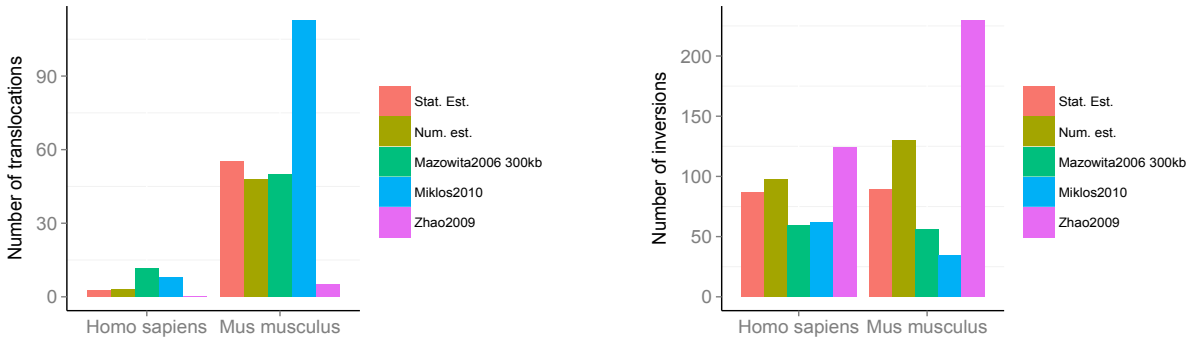
A ratio smaller than one shows that the real G-score is bigger than the simulated, indicating less entropy in the real genome-genome comparisons. Thus, evolution seems to be more systematic than our random model, and improvements should be made to tackle this problem.

The development of this ratio for different inversion size distributions can be found in appendix, Fig. 23. The score stays largely constant, though a small increase can be observed. As the real G-score is fixed, a larger ratio indicates less entropy in our simulated genomes. The seemingly contra-intuitive result of larger inversions reducing the entropy can be solved as follows. Larger inversions can be better detected by PhylDiag, thus it is less necessary to numerically increase the number of inversions. Less inversions however decrease the entropy. However, the effect is small, and an improved version is necessary to use this score in order to measure effects of the inversion size distribution.

7 Discussion

7.1 Comparison with previous estimates

Mazowita et al. (2006) estimate the number of reciprocal translocations and inversions on DNA based on nucleotides. The number of synteny blocks used for this estimation are detected at different resolutions. The higher the resolution, the more synteny blocks they find, thus increasing their estimates. However, they note that at resolutions finer than 100 kb the increase in the estimates is more likely due to noise than due to the fact that more events are observed. If we compare our estimates to their numbers, we most of the time achieve event numbers which are between their 100 kb and 300 kb resolution. As discussed earlier, a human gene with surrounding non-coding DNA has an average size of about 150 kb. Therefore, we



(a) (b)

Figure 12: Comparing our estimations with the literature

Fig. 12(a) compares our estimates of reciprocal translocations on the Homo sapiens - Euarchontoglires and Mus musculus - Euarchontoglires lineage with three other papers. *Stat. est.* is our statistical estimate (table 3, NNLS-MS), and *Num. est.* is our numerical estimate (table 3, Optimal input). Fig. 12(b) shows the same for the number of inversions. #LitAnalysisOfRearrRatesHSMM

can conclude that our estimates are in accordance with their results.

Miklós and Tannier (2010) results are surprising, as they estimate more reciprocal translocations than inversions in the mouse - Euarchontoglires lineage. Additionally, they define another operation, simply called translocations, and estimate nearly the same amount of events for this operation, increasing even more the number of translocations. This is in contrast to other literature. However, they noted in earlier papers that a Markov Chain Monte Carlo method has difficulties to converge on this data. Finally, their models considers only estimates which make it possible to explain the whole sequence of chromosomal events. This constraint may lead to strong deviations from traditional estimates.

The estimations of reciprocal translocations in Zhao and Bourque (2009) are lower, however they introduced an operation called transposition which in our model is equivalent to two reciprocal translocations. Therefore, as they identify many transpositions, consequently, their estimates of reciprocal translocations are lower. However, their estimates of the number of inversions are larger than our estimates and closer to Mazowita et al. (2006) at 100 kb.

7.2 Errors inherent to modelling

In this chapter we discuss different kinds of errors due to simplifications made in our model (chapter 7.2.1), errors based on the nature of phylogenetic trees (chapter 7.2.2) and our biased observation of mutations (chapter 7.2.3)

7.2.1 Simplification of the genome

During modelling, a choice must be about the level of precision in modelling the data. There are different resolutions available, from nucleotide level (every nucleotide is a unit up to chromosome level (several million nucleotides form a unit). All have their applications, and as discussed in chapter 1, there exist already several simulators for different levels. We used modelling in genes for three reasons.

At first, protein coding genes represent a subset of the genome that evolves slowly owing to negative selection. Chromosomal rearrangements that would occur inside a gene would most likely disrupt its function and be counter selected by natural selection. Therefore, they represent markers along chromosomes that are more easily identifiable between species compared to non coding DNA that diverges much more rapidly.

Secondly, the evolutionary history of a gene can be represented in phylogenetic gene trees, which thus provide us with the list of events that occurred along evolution.

Thirdly, protein-coding genes carry essential functions in the genome. Understanding creates therefore much insight about the ancestral animal.

For all these reasons, we decide to treat genes as indivisible markers. Furthermore, the region between two genes, called inter-genic region, has always the same size in our model. For a discussion how to filter out noise in a nucleotide based genome, see Ma et al. (2006) and Ma et al. (2006). This decision has the following three implications. At first, it is not possible that a chromosomal rearrangement happens within the gene. Secondly, there can only be gene events of the whole gene, never only of a part of it. Thirdly, every inter-genic region is treated the same.

To check how much information is lost by this simplifications, we executed several regressions. We want to know if conclusions drawn in our simplified chromosome also hold true in reality. In reality, a reciprocal translocation is an exchange of a chromosome piece of 100,000 bases (100 kb) with another one of 1,000,000 bases (1 mb), where in our model it will be rather 3 genes against 10 genes.

Therefore, we regress chromosome size in base pairs against chromosome size in genes to get an idea of the relationship. We get a $R^2 = 0.84$ in the 5 species used in our numerical optimization (see Fig. 13). This indicates that chromosomes represented by genes can realistically model real chromosome sizes. In appendix, Fig. 19, is a representation of this link for all 21 selected Amniota species. The relationship is in average weaker as for the 5 selected MagSimus species ($R^2 = 0.67$), hence estimations and conclusions based on all 21 species are

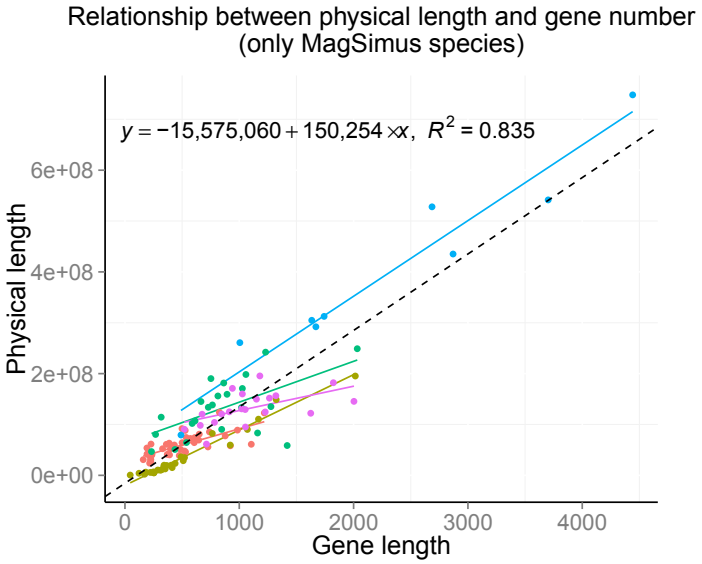


Figure 13: Linear regression of chromosome size in bases on chromosome size in genes in 5 species

The relationship between chromosomes represented in genes (x-axis) and in base pairs (y-axis) is displayed for all 5 individual MagSimus species (coloured solid lines) and for all together (black dashed). As $R^2 = 0.84$, the representation is good and thus we can capture effects based on chromosome size even with our simplified chromosomes rather well. Furthermore, we can conclude that an average gene with surrounding non-coding-sequence has a length of about 150 kb. #ChrSizesInDifferentScalings

more likely error-prone.

Another strong simplification in our model is the treatment of the inter-genic regions. In reality, they may be long, especially fragile or mutation intensive. For example, Berthelot et al. (2015) has shown that the probability for a inter-genic region to break increases with its size in base pairs. Therefore, modelling them all equally is a strong assumption, and may introduce a bias. Furthermore, the probability of a chromosome to undergo a rearrangement will rather change with the amount of non-coding nucleotides (the non-coding size) than with the total nucleotides. Therefore, we regress non-coding chromosome size on chromosome size in genes, as implied by our model. The link is weaker than with chromosome size in total base pairs. However, with $R^2 = 0.79$, the relationship is still rather strong and conclusions on breaking behaviour may be drawn, though cautiously.

7.2.2 Event placing in the tree

We can only observe modern species DNA. That implies that we cannot directly compare the before-after state of DNA to measure the number of mutations. We can only compare the traces of mutations in modern species, look at the placement of these species in the

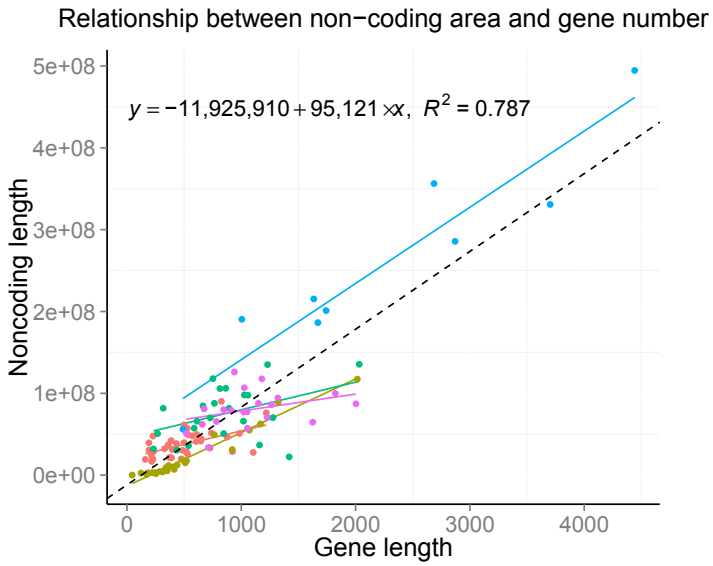


Figure 14: Linear regression of chromosome size in non-coding DNA in bases on chromosome size in genes for 5 species

The relationship between chromosome represented in genes (x-axis) and in non-coding base pairs (y-axis) is displayed for all 5 individual MagSimus species (coloured solid lines) and for all together (black dashed). As $R^2 = 0.79$, the link is weaker than with total size in base pairs. #ChrSizesInDifferentScalings

phylogenetic tree, and with that information deduce the point of occurrence of a mutation.

Yet this process is not error free. An example is given in Fig. 5(a). If we see an event in both species C and D, by parsimony, we would conclude that it happened once between A and B (red circle). Nevertheless, though less likely, it could have happened twice, at the positions of the green triangles. It is not possible to decide in this situation between the two possibilities, thus making the problem non-identifiable.

In reality, one way to partially solve this problem is by looking at other species that are descendants of B. If the type of event is rarely observed, we can think that it is rather unlikely that it happened multiple times on different branches. Yet, it is not unheard of. In 2014, a big group of scientist concluded in a revolutionary series of papers that singing in birds developed not less than 4 times independently, proving that some events may occur more often than we would think by to simple probabilistic calculation. One reason for this may be positive selection of mutations.

Furthermore, events that happen often may be displaced more often. There are several thousand duplications and deletions in our tree. Consider a duplication between A and B gets deleted between B and D. As on some branches in average up to 5% of genes get duplicated and even more deleted, this is not unlikely. Even more, a duplicate of a gene may have

less negative selection of being deleted, and hence the chance of this being observed are even higher. As we observe the duplicate only in C, the duplication event would be wrongly placed between B and C. Therefore, we would expect that we sometimes underestimate the age of gene duplication, and we underestimate the number of gene deletions.

7.2.3 Occurrence rates and observation in modern genomes

Besides in test tube experiments, we cannot observe all mutations that happen. A mutation appears at first only in one individual. By chance or positive selection, the mutation spreads in the population until all new individuals carry the mutation. It is said that the mutation is fixed in the population. However, most mutations are either deleterious or neutral in regard to the fitness level. Deleterious mutations are likely not observed a few generations later. Even if a mutation is neutral, genetic drift may erase the mutation from the gene pool with a certain probability. As we can only observe modern genomes, we only perceive fixed mutations, therefore introducing a bias in the mutations observed. Particularly, it is more accurate to say that we inferred the rates of mutations which got fixed in the population.

Another factor which decreases the number of observation of events is the reversibility of mutations. Though unlikely, we can consider the case of Fig. 5(b). An inversion takes place between A and B, but yet another inversion cancels out the first one between B and D. Hence, instead of counting two inversions, we would count none. At least for bigger chromosomal rearrangements, this case is rather unlikely. This is due to the fact that a chromosomal rearrangement usually involves breaking chromosomes in up to two positions. As a genome has several billion positions to break, or if expressed as a sequence of genes, several thousand positions, breaking twice at the same position (so called breakpoint reuse) is very unlikely. In our simulation, we observed it in less than 1% of the cases, which is comparable to Miklós and Tannier (2010) but much less than Pevzner and Tesler (2003b). Another case of mutation reversibility concerns fissions and fusions, as discussed in chapter 5.2.

8 Conclusions

We provided a probabilistic model of DNA evolution which is based on Mazowita et al. (2006). We showed that it can be used on another, more robust type of data than DNA expressed in nucleotides. Furthermore, we adapted it were the original assumptions were violated and generalized it to include another sort of mutation, gene events. We provided different approaches to calibrate the different parts of the model. Particularly, we provided estimates

for the number of chromosomes in ancestral genomes. Furthermore, we estimated the reciprocal translocation and inversion rates for each branch in a 21 Amniota species phylogeny, based on a more robust dataset than previous estimates. We implemented our model in the gene order simulation MagSimus, created by the group of H. Roest Crollius. This software can now be used to benchmark genome reconstruction softwares. Afterwards, we checked the quality of our estimations and calibration of our model using computer simulations on a sub-sample. We showed that our estimations of translocation rates are robust under the influence of gene events. However, we underestimated the inversion rates on branches with many gene events, particularly gene deletions. This gives us little confidence concerning our inversion rate estimations in birds and ancestral species which lived in the more distant past. We provided numerical estimates for the sub-sample which seem robust to possible biases in our estimation process. Estimates from previous studies show that our estimations based on DNA expressed in genes are comparable to a resolution of 100kb and 300kb for DNA expressed in nucleotides. Subsequently, we analysed the shape of the inversion size distribution in our data. Ultimately, we proposed an entropy score to measure the quality of simulated genomes.

Future improvements in our model and its practical implementation can be made in two aspects: Firstly, intergenic regions should be modelled in more detail. Their size is the main factor for the number of chromosomal rearrangements, and in our recent model we set them all to be equal-sized. This may increase the noise, and thus increasing the size of the confidence intervals of our estimators. Secondly, instead of using random genes to execute gene events, the genes indicated in the gene trees should be selected. This will make the error in the estimation of translocation and inversion distance more realistic.

Future theoretic research may derive an analytical estimator for the reciprocal and inversion distance depending on the influence of noise, in our case gene events. Secondly, our inversion distribution is not optimal. Though most inversions are very small, by looking at genome-genome comparisons we can easily see inversions of several dozen genes. In our model, they have a probability of nearly 0, therefore we should modify our inversion distribution to have a larger tail. As discussed in this thesis, an Bayesian analysis with a wider range of possible distributions and a modified fitness statistic may provide a direct approach to improve upon our results.

References

Alekseyev, M. and P. Pevzner (2009): "Breakpoint graphs and ancestral genome reconstructions." *Genome research*.

Batut, B., D. Parsons, S. Fischer, G. Beslon, and C. Knibbe (2013): "In silico experimental evolution: a tool to test evolutionary scenarios," *BMC Bioinformatics*, 14, S11.

Bérard, S., C. Gallien, B. Boussau, G. Szllosi, V. Daubin, and E. Tannier (2012): "Evolution of gene neighborhoods within reconciled phylogenies," *Bioinformatics*, 28, 382–388.

Berthelot, C., M. Muffato, J. Abecassis, and H. RoestCrollius (2015): "The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions," *Cell Reports*, 10, 1913–1924.

Chauve, C., N. E. Mabrouk, and E. Tannier (2013): *Models and Algorithms for Genome Evolution*, Springer Publishing Company, Incorporated.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girn, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Khri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek (2015): "Ensembl 2015," *Nucleic Acids Research*, 43, D662–D669.

Dalquen, D. A., M. Anisimova, G. H. Gonnet, and C. Dessimoz (2012): "ALFA Simulation Framework for Genome Evolution," *Molecular Biology and Evolution*, 29, 1115–1123.

De, A., M. Ferguson, S. Sindi, and R. Durrett (2001): "The equilibrium distribution for a generalized Sankoff-Ferretti model accurately predicts chromosome size distributions in a wide variety of species," *J. Appl. Probab.*, 38, 324–334.

Deakin, J. and T. Ezaz (2014): "Tracing the evolution of amniote chromosomes," *Chromosoma*, 123, 201–216.

Felsenstein, J. (2004): *Inferring Phylogenies*, Sunderland, Massachusetts: Sinauer Associates, Inc., 5 ed.

Ferretti, V., J. H. Nadeau, and D. Sankoff (1996): "Original Synteny," in *Proceedings of the 7th Annual Symposium on Combinatorial Pattern Matching*, London, UK, UK: Springer-Verlag, CPM '96, 159–167.

Gascuel, O. (2007): *Mathematics of Evolution and Phylogeny*, New York, NY, USA: Oxford University Press, Inc.

Glick, L. and I. Mayrose (2014): "ChromEvol: Assessing the Pattern of Chromosome Number Evolution and the Inference of Polyploidy along a Phylogeny," *Mol Biol Evol*, 31, 1914–1922.

Gregory, T. (2015): "Animal Genome Size Database," .

Hedges, S., J. Marin, M. Suleski, M. Paymer, and S. Kumar (2015): "Tree of Life Reveals Clock-Like Speciation and Diversification," *Mol Biol Evol*, 32, 835–845.

Hoey, J. (2012): "The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test," .

Huttley, G. A., M. J. Wakefield, and S. Easteal (2007): "Rates of Genome Evolution and Branching Order from Whole Genome Analysis," *Mol Biol Evol*, 24, 1722–1730.

Lin, Y., V. Rajan, K. Swenson, and B. Moret (2010): "Estimating true evolutionary distances under rearrangements, duplications, and losses," *BMC Bioinformatics*, 11.

Lucas, J., M. Muffato, and H. Crollius (2014): "PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees," *BMC Bioinformatics*, 15, 268+.

Ma, J., L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, and W. Miller (2006): "Reconstructing contiguous regions of an ancestral genome," *Genome Research*, 16, 1557–1565.

Mayrose, I., M. Barker, and O. S.P. (2010): "Probabilistic Models of Chromosome Number Evolution and the Inference of Polyploidy," *Mol Biol Evol*, 59, 132–144.

Mazowita, M., L. Haque, and D. Sankoff (2006): "Stability of rearrangement measures in the comparison of genome sequences." *J Comput Biol*, 13, 554–566.

Miklós, I. and E. Tannier (2010): "Bayesian sampling of genomic rearrangement scenarios via double cut and join," *Bioinformatics*, 26, 3012–3019.

Muffato, M. (2010): "Reconstruction de génomes ancestraux chez les vertébrés," Dissertation, Universit d'Evry-Val d'Essonne.

Mullen, K. M. and I. H. M. van Stokkum (2012): *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*, r package version 1.4.

Nelson, C. and S. Vialette, eds. (2008): *Comparative Genomics*, Berlin: Springer-Verlag, 1 ed.

Ouangraoua, A., E. Tannier, and C. Chauve (2011): "Reconstructing the architecture of the ancestral amniote genome." *Bioinformatics*, 27, 2664–2671.

Paten, B., J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, I. Holmes, and E. Birney (2008): "Genome-wide nucleotide-level mammalian ancestor reconstruction," *Genome research*, 18, 1829–1843.

Peres, A. and H. Roest Crollius (2015): "Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes," *BMC Bioinformatics*, 16(Suppl 3):A9.

Pevzner, P. and G. Tesler (2003a): "Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes," *Genome Research*, 13, 37–45.

——— (2003b): "Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution," *Proc Natl Acad Sci U S A*, 100, 7672–7.

——— (2003c): "Transforming Men into Mice: The Nadeau-Taylor Chromosomal Breakage Model Revisited," in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, New York, NY, USA: ACM, RECOMB '03, 247–256.

Reddy, T.B.K.., A. Thomas, D. Stamatis, J. Bertsch, M. Isbandi, J. Jansson, J. Mallajosyula, I. Pagani, E. Lobos, and N. Kyrpides (2015): "The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification," *Nucl Acids Res*, 43, 1099–1106.

Sankoff, D. and M. Mazowita (2005): "Stability of Rearrangement Measures in the Comparison of Genome Sequences," in *Research in Computational Molecular Biology, 9th Annual International Conference, RECOMB 2005, Cambridge, MA, USA, May 14-18, 2005, Proceedings*, 603–614.

Zhao, H. and G. Bourque (2009): "Recovering genome rearrangements in the mammalian phylogeny." *Genome research*, 19, 934–942.

A Figures

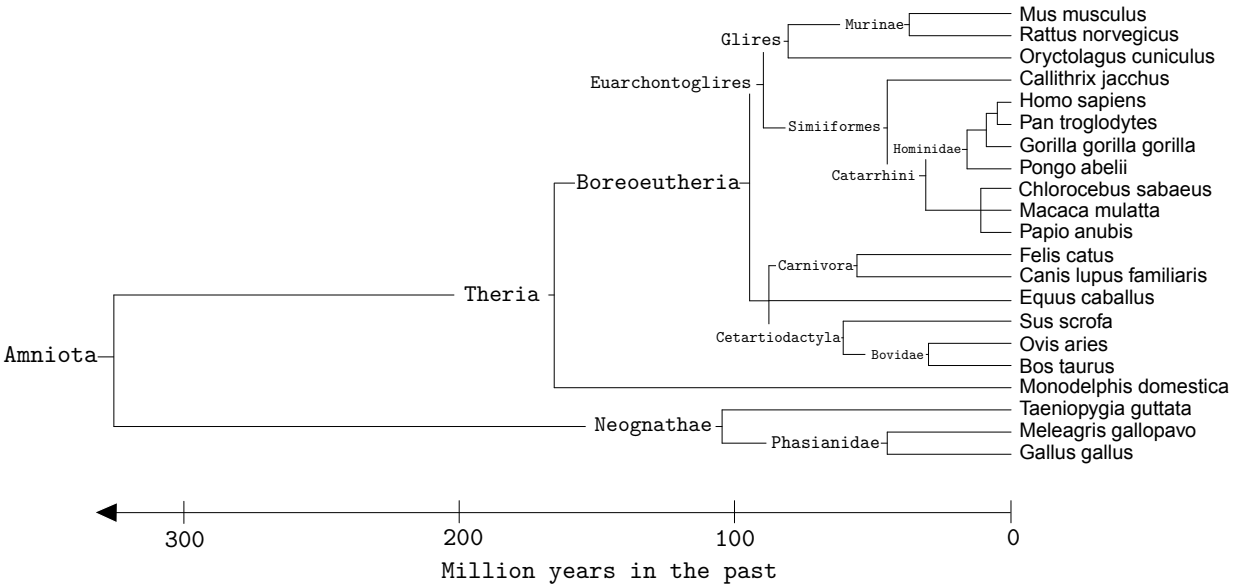


Figure 15: Phylogenetic tree with branch length according to Ensembl 78 for all selected Amniota species

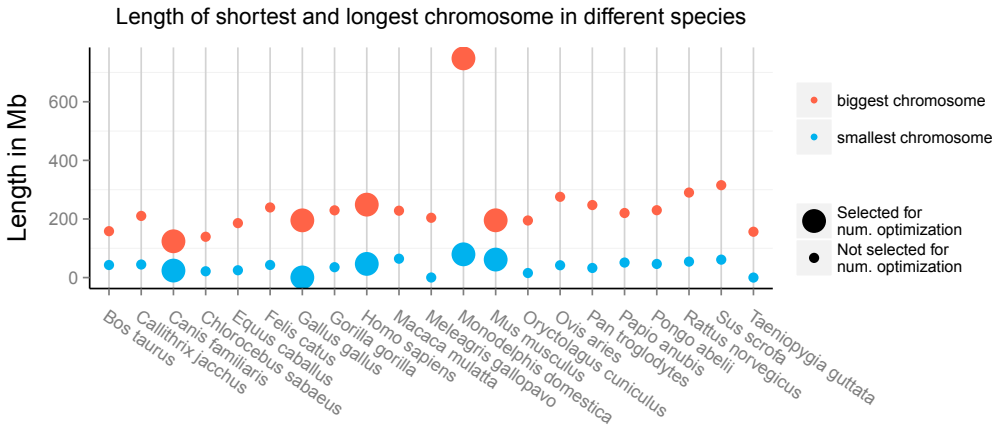


Figure 16: Minimal and maximal chromosome sizes in selected Amniota species in base pairs. The upper, red dots indicate the size of the largest chromosome of a species, the lower blue dots indicate the size of the smallest chromosome in a species, not including Y and W chromosomes. The larger dots indicate species which were selected to be simulated with the genome simulator MagSimus. As can be seen, there is large variation in maximal and minimal chromosome size, and hence no conclusion can be drawn on physical constraints on upper or lower chromosome sizes. There are both, very large macro-chromosomes in Monodelphis domestica as well as micro-chromosomes in birds as Gallus gallus. #ChrSizesMinMaxAmniota

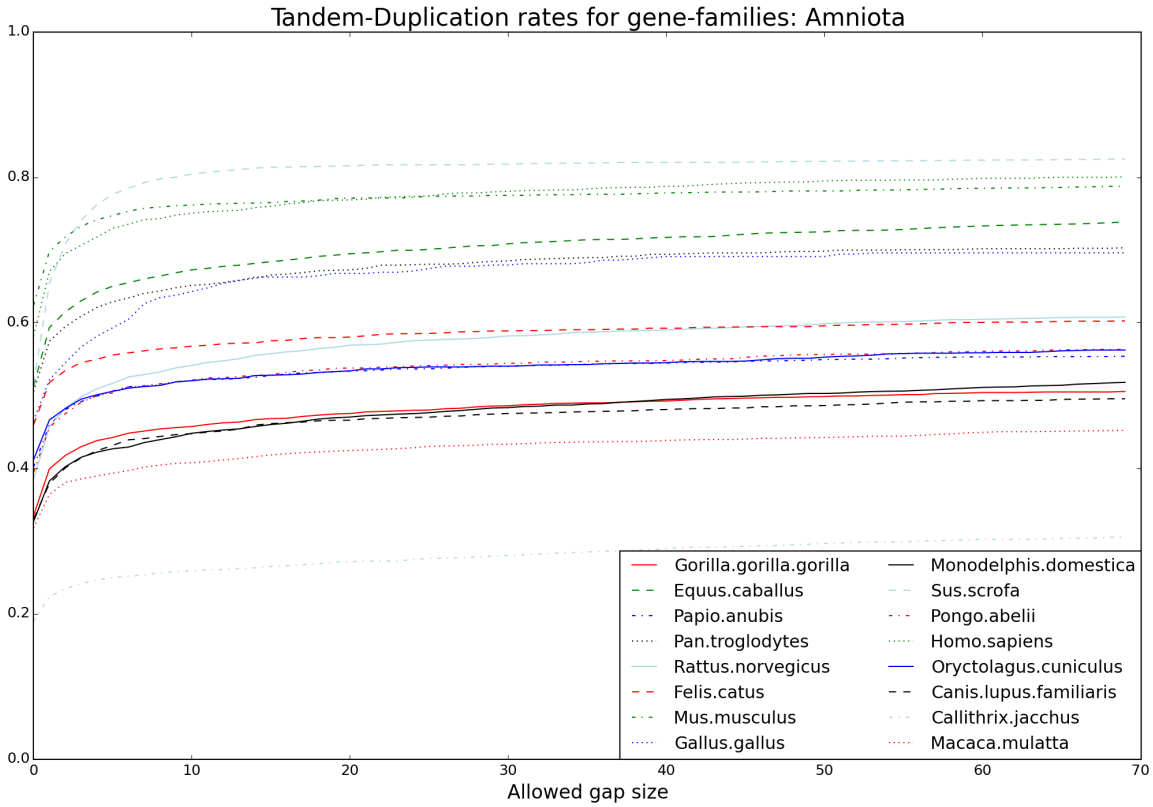


Figure 17: Cumulative density function for duplication distances

The proportion of close to distant duplications varies heavily between different species. However, all distributiosn have in common that after a rapid decrease, they stay nearly constant after gap 10, indicating the existence of two types of duplications: short distance duplications, where appearance depends heavily on the distance to the original gene, and distant duplications, which appear uniformly over the genome.

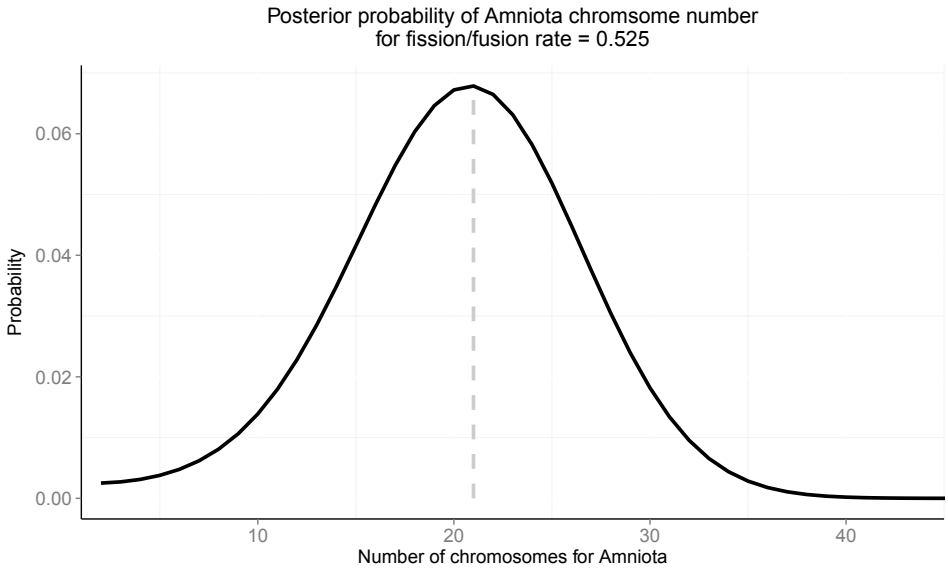


Figure 18: Probability density for different start chromosome number for Amniota genome as calculated by ChromEvol 2

The most likely number of chromosomes for the Amniota genome was estimated to be 21. #ProbAmniotaChromosomeNumber

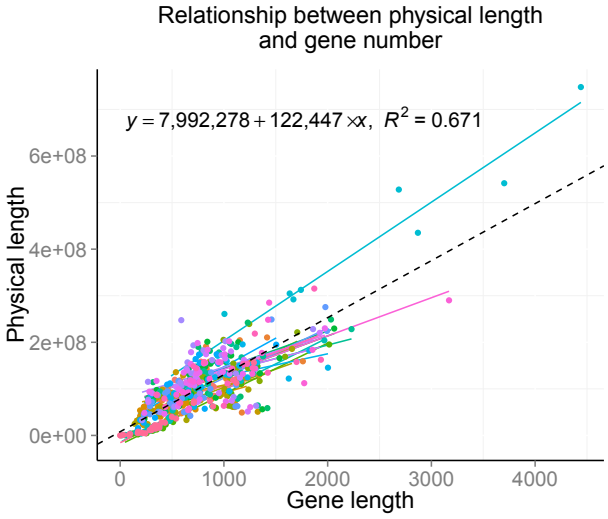


Figure 19: Linear regression of chromosome size in bases on chromosome size in genes

The relationship between chromosome represented in genes (x-axis) and in bases (y-axis) is displayed for all 5 species selected for numerical optimization (coloured solid lines) and for all together (black dashed). As $R^2 = 0.83$, the representation is good and thus we conclude that we can capture effects based on chromosome size even with our simplified chromosomes rather well. #ChrSizesInDifferentScalings

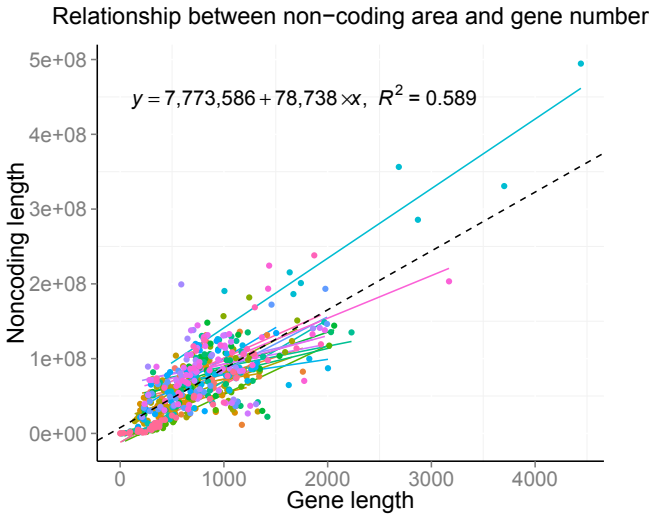


Figure 20: The relationship between chromosomes represented in genes (x-axis) and in non-coding bases (y-axis) is displayed for all selected Amniota species (coloured solid lines) and for all together (black dashed). As $R^2 = 0.59$, the link is a lot weaker than in the 5 species selected for numerical optimization alone (Fig. 14). A positive correlation indicates that more genes on a chromosome also indicate more non-coding DNA, i.e. more space for potential chromosome rearrangements. #ChrSizesInDifferentScalings

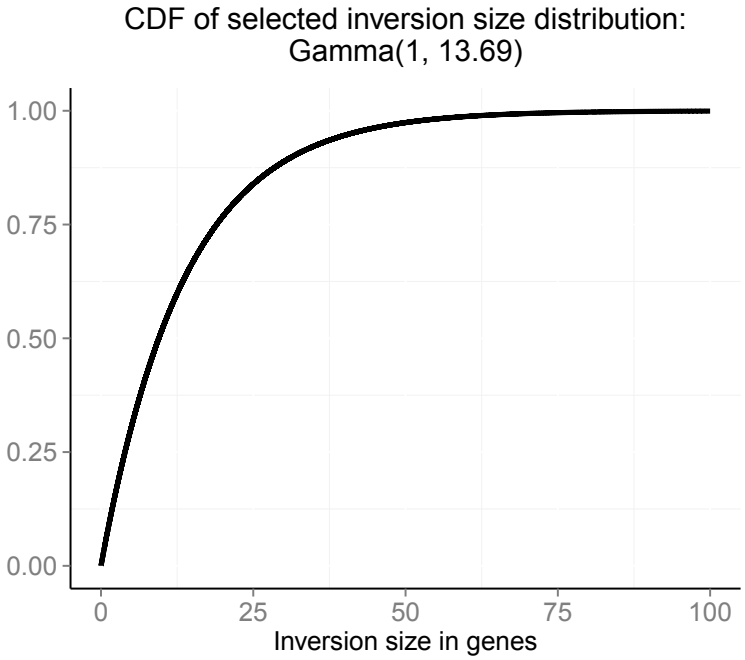
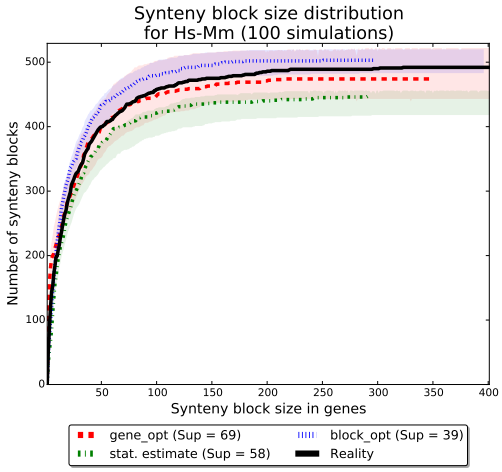
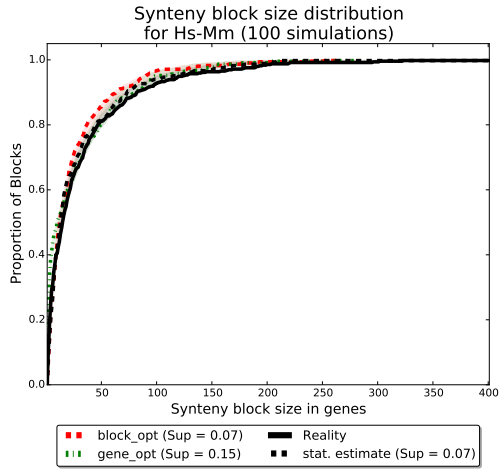


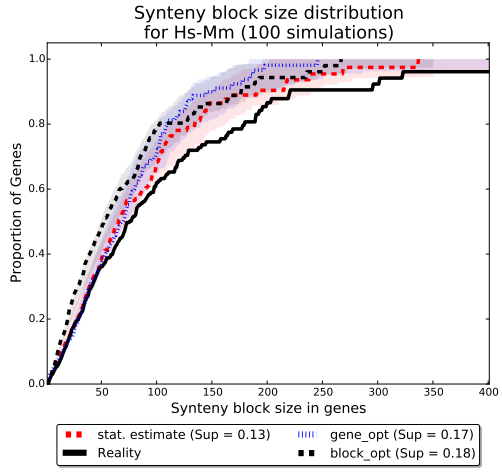
Figure 21: Cumulative density function for proposed inversion size distribution, where the size of the inversion is measured in genes. #InversionSizeDistribution



(a) Ordinate scaled to number of synteny blocks



(b) Ordinate scaled to proportion of synteny blocks



(c) Ordinate scaled to proportion of genes

Figure 22: Comparison of fits to synteny block size distribution

The plots shows different synteny block size distributions for different ordinate scales. Distribution in real values (*Reality*), simulated with our statistical estimates (*stat. estimat*), simulated with values for optimal Kolmogorov-Smirnoff statistic by synteny blocks (*block opt*), and simulated with values for Kolmogorov-Smirnoff statistic by genes (*gene opt*) for different ordinate scales. The maximal distance to reality is indicated in parenthesis. The coloured area indicates 95% of simulations.

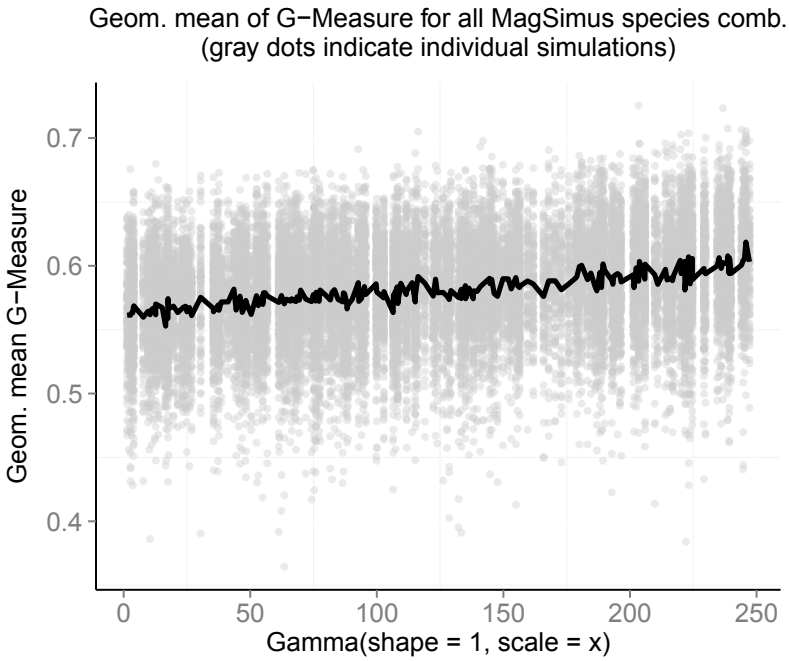


Figure 23: G-Score development for different inversion size distributions

The black line indicates the mean of 100 replications, indicated as gray dots, per scale parameter. The score stays largely constant, though a small increase can be observed. See chapter 6.3 for a discussion. #InversionSizeDistribution

B Tables

Input	Question	Answer
Fission	Rate ?	Determined by a combination of ChromEvol 2 and parsominous
	Which chromosome should break ?	Random chromosome is sampled, independently of size
	Where to break on chromosome ?	Randomly sampled position strictly within chromosome
Fusion	Rate ?	Determined by a combination of ChromEvol 2 and parsominous
	Which 2 chromosomes should fuse ?	2 random chromosomes sampled without replacement, independently of size
	Which endings combine?	For each chromosome, randomly sample one of both endings
Reciprocal translocation	Rate?	Determined with PhylDiag, Mazowita 2006 estimator and Non-negative least squares estimation
	Which 2 chromosomes are involved ?	2 random chromosomes sampled without replacement, independently of size
	Where to break both chromosomes ?	2 randomly sampled positions strictly within chromosome
	Which endings combine?	Combine 2 random inner endings
Inversion	Rate?	Determined with PhylDiag, Mazowita 2006 estimator and Non-negative least squares estimation
	Where to break chromosome ? Or: How long should inversion be ?	Size of inversion is sampled by Gamma(1, 13.69), random breakpoint is sampled such that inversion of sampled size can take place strictly within chromosome
	Which chromosome is involved ?	Random chromosome is sampled among those which are large enough for sampled inversion
Gene duplication	Rate ?	MagSimus 1: Naïve rate inferred from improved Ensembl 78 gene trees MagSimus 2: Strictly follows improved Ensembl 78 gene trees
	Which gene should duplicate ?	MagSimus 1: Random gene chosen MagSimus 2: Strictly follows improved Ensembl 78 gene trees
	Where to place duplication ? Or: Ratio of tandem duplications ?	Simple model: either gene is duplicated side by side (tandem duplication) or gene is randomly placed in genome (distant duplication); Tandem duplication ratio was inferred on Ensembl 78 genomes
	Direction of duplication ?	In average, roughly 75% of duplicates in modern genomes same direction, hence 75% was taken
Gene deletion	Rate ?	MagSimus 1: Naïve rate inferred from improved Ensembl 78 gene trees MagSimus 2: Strictly follows improved Ensembl 78 gene trees
	Which gene should be deleted ?	MagSimus 1: Random gene chosen MagSimus 2: Strictly follows improved Ensembl 78 gene trees
Start genome	Number of chromosomes ?	Determined by ChromEvol 2
	Sizes of chromosomes ?	Gene number determined by linear regression on Ensembl 78 Amniota genomes, Chromosome size distribution inferred by averaging over interpolated Ensembl 78 chromosome size distributions

Table 4: Summary of model choices.

Amniota species in Ensemble 78	Genes	Genes deleted	Not deleted %	> 50% of genes preserved
Ailuropoda melanoleuca	19343	19343	0.00	No
Anas platyrhynchos	15634	15634	0.00	No
Anolis carolinensis	18596	9528	48.76	No
Bos taurus	19994	26	99.87	Yes
Callithrix jacchus	20978	937	95.53	Yes
Canis lupus familiaris	19856	282	98.58	Yes
Cavia porcellus	18673	18673	0.00	No
Chlorocebus sabaeus	19165	252	98.69	Yes
Choloepus hoffmanni	12393	12393	0.00	No
Dasypus novemcinctus	22711	22711	0.00	No
Dipodomys ordii	15798	15798	0.00	No
Echinops telfairi	16575	16575	0.00	No
Equus caballus	20449	192	99.06	Yes
Erinaceus europaeus	14601	14601	0.00	No
Felis catus	19493	314	98.39	Yes
Ficedula albicollis	15303	15303	0.00	No
Gallus gallus	15508	1049	93.24	Yes
Gorilla gorilla gorilla	20962	42	99.80	Yes
Homo sapiens	21796	2254	89.66	Yes
Ictidomys tridecemlineatus	18826	18826	0.00	No
Loxodonta africana	20033	20033	0.00	No
Macaca mulatta	21905	882	95.97	Yes
Macropus eugenii	15290	15290	0.00	No
Meleagris gallopavo	14123	832	94.11	Yes
Microcebus murinus	16319	16319	0.00	No
Monodelphis domestica	21327	1087	94.90	Yes
Mus musculus	22154	360	98.38	Yes
Mustela putorius furo	19910	19910	0.00	No
Myotis lucifugus	19728	19728	0.00	No
Nomascus leucogenys	18575	18575	0.00	No
Ochotona princeps	16006	16006	0.00	No
Ornithorhynchus anatinus	21698	18996	12.45	No
Oryctolagus cuniculus	19293	5393	72.05	Yes
Otolemur garnettii	19506	19506	0.00	No
Ovis aries	20921	784	96.25	Yes
Pan troglodytes	18759	580	96.91	Yes
Papio anubis	19210	618	96.78	Yes
Pelodiscus sinensis	18189	18189	0.00	No
Pongo abelii	20424	1306	93.61	Yes
Procavia capensis	16057	16057	0.00	No
Pteropus vampyrus	16990	16990	0.00	No
Rattus norvegicus	22776	76	99.67	Yes
Sarcophilus harrisii	18788	18788	0.00	No
Sorex araneus	13187	13187	0.00	No
Sus scrofa	21605	2189	89.87	Yes
Taeniopygia guttata	17488	3919	77.59	Yes
Tarsius syrichta	13628	13628	0.00	No
Tupaia belangeri	15471	15471	0.00	No
Tursiops truncatus	16550	16550	0.00	No
Vicugna pacos	11765	11765	0.00	No

Table 5: Amniota species in Ensembl 78 with their respective number of coding genes

During the cleansing process, both genes on mitochondrial DNA as well as on scaffolds were excluded. If after this cleansing more than 50% of the genes were preserved, the species was included in the database.

	MagSimus branches	Rate per mya	Number of events
Gene Birth	Boreoeutheria	18.35	1303
	Canis lupus familiaris	6.78	644
	Euarchontoglires	26.00	130
	Gallus gallus	2.59	843
	Homo sapiens	4.71	424
	Monodelphis domestica	5.57	925
	Mus musculus	13.47	1212
	Theria	5.35	856
Gene Duplications	Boreoeutheria	30.04	2133
	Canis lupus familiaris	14.08	1338
	Euarchontoglires	73.80	369
	Gallus gallus	1.81	590
	Homo sapiens	13.56	1220
	Monodelphis domestica	20.98	3482
	Mus musculus	30.96	2786
	Theria	7.44	1191
Gene Loss	Boreoeutheria	10.90	774
	Canis lupus familiaris	61.61	5853
	Euarchontoglires	254.40	1272
	Gallus gallus	22.29	7266
	Homo sapiens	53.04	4774
	Monodelphis domestica	29.82	4950
	Mus musculus	54.18	4876
	Theria	9.73	1556
Tandem Duplication Proportion	Boreoeutheria	0.614	
	Canis lupus familiaris	0.348	
	Euarchontoglires	0.542	
	Gallus gallus	0.651	
	Homo sapiens	0.798	
	Monodelphis domestica	0.443	
	Mus musculus	0.822	
	Theria	0.652	

Table 6: Branch estimates for different gene events for a phylogenetic tree of 5 Amniota species.

Selected Amniota branches	SB	Translocations				Inversions			
	NNLS	LM	LM Weighted	NNLS	NNLS Unif	LM	LM Weighted	NNLS	NNLS Unif
Boreoeutheria	138.71	19.54	19.41	19.18	16.51	50.24	51.21	50.64	53.12
Bos taurus	225.42	0.75	0.88	0.75	1.10	103.93	105.37	103.93	103.59
Bovidae	141.81	7.82	7.69	7.82	7.89	86.87	71.15	60.25	60.30
Callithrix jacchus	301.10	10.40	10.70	10.40	10.12	134.16	136.25	134.16	134.44
Canis lupus familiaris	205.68	21.36	21.10	21.82	21.10	68.20	69.80	68.20	68.69
Carnivora	52.64	2.26	2.26	1.75	2.06	30.60	22.69	23.86	23.68
Catarrhini	53.50	1.05	0.75	0.96	1.19	25.92	23.89	25.92	25.71
Cercopithecinae	54.40	1.30	1.23	1.34	1.33	26.10	26.58	26.10	26.10
Cetartiodactyla	0.00	1.84	1.96	1.84	1.54	-22.54	-7.14	0.00	0.00
Chlorocebus sabaeus	33.92	1.32	1.36	1.32	1.45	7.18	15.84	7.18	7.04
Equus caballus	176.04	6.07	5.94	6.07	6.27	78.00	72.36	71.97	71.65
Euarchontoglires	17.55	-1.39	-0.73	0.00	0.00	14.16	12.35	8.91	9.58
Felis catus	96.32	-0.56	-0.30	0.00	0.00	42.50	40.90	42.50	42.20
Gallus gallus	102.11	-0.92	-0.65	0.00	0.00	44.62	43.63	44.62	44.02
Glires	8.47	1.96	2.48	1.74	0.98	2.01	5.86	2.01	2.65
Gorilla gorilla gorilla	103.77	0.98	1.07	1.10	1.15	41.43	40.86	44.70	44.66
Hominidae	30.25	-0.29	-0.13	0.00	0.00	15.03	14.33	15.03	15.00
Homininae	36.37	0.64	0.48	0.27	0.24	21.89	22.12	17.75	17.79
Homo sapiens	34.82	0.67	0.62	0.37	0.38	19.06	15.25	11.20	11.19
HomoPan	0.00	-0.47	-0.56	0.00	0.00	-12.44	-11.87	0.00	0.00
Laurasiatheria	0.00	2.01	1.55	1.49	1.49	-13.31	-11.50	0.00	0.00
Macaca mulatta	138.76	2.22	2.07	2.22	2.15	61.27	53.38	61.27	61.35
Meleagris gallopavo	243.89	12.40	12.12	12.38	11.21	101.40	102.40	101.40	102.15
Monodelphis domestica	456.56	8.37	8.37	8.37	6.62	213.90	213.90	213.90	215.64
Murinae	95.14	42.37	42.28	42.37	38.68	5.04	3.46	5.04	8.73
Mus musculus	198.58	8.93	9.26	8.93	9.78	85.15	81.34	85.15	84.30
Neognathae	191.57	4.31	4.31	3.07	3.56	90.63	90.63	90.63	90.25
Oryctolagus cuniculus	90.86	1.77	1.86	1.77	2.50	38.07	39.65	38.07	37.34
Ovis aries	167.58	2.30	2.17	2.30	1.97	74.52	73.08	74.52	74.84
Pan troglodytes	36.40	1.38	1.42	1.08	1.04	18.90	22.70	11.04	11.08
Papio anubis	46.82	0.06	0.17	0.06	0.01	17.46	16.68	17.46	17.51
Phasianidae	103.04	-1.90	-1.90	0.00	0.00	53.58	53.58	53.58	53.89
Pongo abelii	48.96	0.02	0.22	0.00	0.00	18.41	17.99	18.41	18.37
Rattus norvegicus	581.42	11.87	11.54	11.87	10.68	273.55	277.36	273.55	274.74
Simiiformes	56.48	3.99	3.72	3.32	3.66	24.27	22.52	24.27	23.57
Sus scrofa	1086.68	11.27	11.40	11.27	10.86	548.04	563.76	524.22	524.73
Taeniopygia guttata	181.46	1.56	1.56	3.24	3.45	80.25	80.25	80.25	79.90
Theria	138.69	3.12	3.12	2.23	2.58	65.62	65.62	65.62	65.34

Table 7: Estimated number of reciprocal translocations and inversions on branches of phylogenetic tree Fig. 15

NNLS: Non-negative least square estimation. LM: Linear model. LM Weighted: Linear model with weights being indirectly proportional to phylogenetic distances. NNLS Unif: Non-negative least square estimation based onreciprocal translocation and inversion distances created by a modified Mazowita et al. (2006), equation 9. See chapter 2.4.2.

Selected Amniota species combinations	SB	Inv	Inv Unif	Transl	Transl Unif
Bos taurus - Callithrix jacchus	763	343.0	343.2	23.5	23.3
Bos taurus - Canis lupus familiaris	604	242.1	242.7	40.4	39.8
Bos taurus - Chlorocebus sabaeus	552	240.6	240.7	20.4	20.3
Bos taurus - Equus caballus	531	230.2	230.3	19.3	19.2
Bos taurus - Felis catus	500	225.6	225.6	9.4	9.4
Bos taurus - Gallus gallus	1031	460.9	462.2	39.6	38.3
Bos taurus - Gorilla gorilla gorilla	673	304.7	304.8	16.8	16.7
Bos taurus - Homo sapiens	585	262.3	262.3	15.2	15.2
Bos taurus - Macaca mulatta	669	302.5	302.6	17.0	16.9
Bos taurus - Meleagris gallopavo	1183	523.7	526.1	52.3	49.9
Bos taurus - Monodelphis domestica	948	428.5	429.9	30.5	29.1
Bos taurus - Mus musculus	701	268.7	270.9	66.8	64.6
Bos taurus - Oryctolagus cuniculus	482	211.3	211.3	14.7	14.7
Bos taurus - Ovis aries	393	178.4	178.4	3.1	3.1
Bos taurus - Pan troglodytes	584	260.7	260.8	16.3	16.2
Bos taurus - Papio anubis	587	263.7	263.8	14.8	14.7
Bos taurus - Pongo abelii	563	252.4	252.4	14.1	14.1
Bos taurus - Rattus norvegicus	1108	467.2	471.0	71.8	68.0
Bos taurus - Sus scrofa	1555	742.0	742.3	20.5	20.2
Bos taurus - Taeniopygia guttata	1008	446.1	447.3	41.4	40.2
Callithrix jacchus - Canis lupus familiaris	646	259.5	260.4	44.0	43.1
Callithrix jacchus - Chlorocebus sabaeus	432	185.3	185.4	15.7	15.6
Callithrix jacchus - Equus caballus	517	216.3	216.7	26.2	25.8
Callithrix jacchus - Felis catus	502	223.4	223.5	16.1	16.0
Callithrix jacchus - Gallus gallus	1036	467.1	468.6	36.4	34.9
Callithrix jacchus - Gorilla gorilla gorilla	553	251.3	251.4	13.2	13.1
Callithrix jacchus - Homo sapiens	415	184.4	184.5	11.6	11.5
Callithrix jacchus - Macaca mulatta	597	272.1	272.2	14.9	14.8
Callithrix jacchus - Meleagris gallopavo	1195	530.4	533.6	51.6	48.4
Callithrix jacchus - Monodelphis domestica	940	423.4	426.1	35.1	32.4
Callithrix jacchus - Mus musculus	629	233.5	236.6	69.5	66.4
Callithrix jacchus - Oryctolagus cuniculus	465	205.5	205.7	15.5	15.3
Callithrix jacchus - Ovis aries	668	295.7	296.1	24.8	24.4
Callithrix jacchus - Pan troglodytes	422	186.9	187.0	12.1	12.0
Callithrix jacchus - Papio anubis	459	205.8	205.9	12.2	12.1
Callithrix jacchus - Rattus norvegicus	1064	444.3	449.9	76.2	70.6
Callithrix jacchus - Sus scrofa	1557	742.0	742.5	25.0	24.5
Callithrix jacchus - Taeniopygia guttata	1011	449.6	451.1	39.4	37.9
Callithrix jacchus - Pongo abelii	441	197.5	197.6	11.0	10.9
Canis lupus familiaris - Chlorocebus sabaeus	430	163.2	163.6	32.3	31.9
Canis lupus familiaris - Equus caballus	381	138.2	138.7	32.8	32.3
Canis lupus familiaris - Felis catus	302	110.7	110.9	20.8	20.6
Canis lupus familiaris - Gallus gallus	1002	426.3	428.7	55.2	52.8
Canis lupus familiaris - Gorilla gorilla gorilla	535	219.2	219.6	28.8	28.4
Canis lupus familiaris - Homo sapiens	434	171.4	171.7	26.1	25.8
Canis lupus familiaris - Macaca mulatta	550	225.3	225.7	30.2	29.8
Canis lupus familiaris - Meleagris gallopavo	1142	482.2	486.0	69.3	65.5
Canis lupus familiaris - Monodelphis domestica	880	375.4	378.4	45.1	42.1
Canis lupus familiaris - Mus musculus	564	178.0	181.6	84.5	80.9
Canis lupus familiaris - Oryctolagus cuniculus	401	155.9	156.2	25.1	24.8
Canis lupus familiaris - Ovis aries	546	219.8	220.4	33.7	33.1

Canis lupus familiaris - Pan troglodytes	441	174.4	174.7	26.6	26.3
Canis lupus familiaris - Papio anubis	441	174.2	174.5	26.8	26.5
Canis lupus familiaris - Pongo abelii	418	165.7	165.9	23.8	23.6
Canis lupus familiaris - Rattus norvegicus	977	384.3	389.3	84.7	79.7
Canis lupus familiaris - Sus scrofa	1462	679.8	680.4	31.7	31.1
Canis lupus familiaris - Taeniopygia guttata	973	411.5	413.5	55.5	53.5
Chlorocebus sabaeus - Equus caballus	337	135.9	136.0	16.6	16.5
Chlorocebus sabaeus - Felis catus	279	116.1	116.2	8.4	8.3
Chlorocebus sabaeus - Gallus gallus	912	407.2	408.2	33.8	32.8
Chlorocebus sabaeus - Gorilla gorilla gorilla	265	114.9	114.9	2.6	2.6
Chlorocebus sabaeus - Homo sapiens	201	84.0	84.0	1.5	1.5
Chlorocebus sabaeus - Macaca mulatta	184	73.4	73.4	3.6	3.6
Chlorocebus sabaeus - Meleagris gallopavo	1068	470.4	472.6	48.1	45.9
Chlorocebus sabaeus - Monodelphis domestica	790	350.9	352.2	29.1	27.8
Chlorocebus sabaeus - Mus musculus	453	152.5	154.5	59.0	57.0
Chlorocebus sabaeus - Oryctolagus cuniculus	329	138.0	138.0	11.5	11.5
Chlorocebus sabaeus - Ovis aries	476	202.9	203.1	20.1	19.9
Chlorocebus sabaeus - Pan troglodytes	220	91.9	91.9	3.1	3.1
Chlorocebus sabaeus - Papio anubis	141	52.9	52.9	2.6	2.6
Chlorocebus sabaeus - Pongo abelii	162	64.5	64.5	1.5	1.5
Chlorocebus sabaeus - Rattus norvegicus	870	358.4	361.5	61.6	58.5
Chlorocebus sabaeus - Sus scrofa	1425	675.9	676.1	21.6	21.4
Chlorocebus sabaeus - Taeniopygia guttata	900	398.4	399.3	35.1	34.2
Equus caballus - Felis catus	262	109.8	109.8	5.2	5.2
Equus caballus - Gallus gallus	885	392.1	393.3	34.4	33.2
Equus caballus - Gorilla gorilla gorilla	420	181.5	181.5	12.5	12.5
Equus caballus - Homo sapiens	344	144.0	144.0	12.0	12.0
Equus caballus - Macaca mulatta	481	210.8	210.9	13.7	13.6
Equus caballus - Meleagris gallopavo	1044	457.3	459.6	48.7	46.4
Equus caballus - Monodelphis domestica	895	397.9	400.0	33.6	31.5
Equus caballus - Mus musculus	482	166.7	168.9	58.3	56.1
Equus caballus - Oryctolagus cuniculus	373	160.6	160.6	9.9	9.9
Equus caballus - Ovis aries	463	199.3	199.4	16.2	16.1
Equus caballus - Pan troglodytes	378	159.4	159.5	13.6	13.5
Equus caballus - Papio anubis	345	144.4	144.5	12.1	12.0
Equus caballus - Pongo abelii	327	136.6	136.6	10.9	10.9
Equus caballus - Rattus norvegicus	908	380.5	383.5	57.5	54.5
Equus caballus - Sus scrofa	1414	673.3	673.5	17.7	17.5
Equus caballus - Taeniopygia guttata	884	388.1	389.2	37.4	36.3
Felis catus - Gallus gallus	921	417.2	418.2	28.8	27.8
Felis catus - Gorilla gorilla gorilla	403	179.0	179.0	10.5	10.5
Felis catus - Homo sapiens	306	131.5	131.6	10.0	9.9
Felis catus - Macaca mulatta	427	190.2	190.3	12.8	12.7
Felis catus - Meleagris gallopavo	1090	487.6	489.8	41.9	39.7
Felis catus - Monodelphis domestica	799	366.4	367.5	23.6	22.5
Felis catus - Mus musculus	427	148.7	150.4	54.8	53.1
Felis catus - Oryctolagus cuniculus	314	136.5	136.6	9.5	9.4
Felis catus - Ovis aries	448	196.2	196.4	14.3	14.1
Felis catus - Pan troglodytes	325	140.0	140.0	10.5	10.5
Felis catus - Papio anubis	311	134.4	134.5	10.6	10.5
Felis catus - Pongo abelii	290	124.1	124.1	8.9	8.9
Felis catus - Rattus norvegicus	865	364.8	368.0	57.2	54.0
Felis catus - Sus scrofa	1376	661.0	661.2	17.5	17.3
Felis catus - Taeniopygia guttata	904	405.4	406.2	30.1	29.3
Gallus gallus - Gorilla gorilla gorilla	1001	455.4	456.7	30.6	29.3

Gallus gallus - Homo sapiens	937	424.5	425.7	29.5	28.3
Gallus gallus - Macaca mulatta	1033	470.4	471.8	31.6	30.2
Gallus gallus - Meleagris gallopavo	346	146.0	146.2	11.5	11.3
Gallus gallus - Monodelphis domestica	1005	476.5	476.9	11.5	11.1
Gallus gallus - Mus musculus	1023	428.5	434.7	68.5	62.3
Gallus gallus - Oryctolagus cuniculus	755	339.9	340.7	23.1	22.3
Gallus gallus - Ovis aries	997	445.0	447.1	39.0	36.9
Gallus gallus - Pan troglodytes	918	415.1	416.3	29.4	28.2
Gallus gallus - Papio anubis	942	427.4	428.6	29.1	27.9
Gallus gallus - Pongo abelii	914	414.4	415.4	28.1	27.1
Gallus gallus - Rattus norvegicus	1340	582.1	590.5	73.4	65.0
Gallus gallus - Sus scrofa	1404	652.1	654.2	35.4	33.3
Gallus gallus - Taeniopygia guttata	373	168.5	168.5	1.5	1.5
Gorilla gorilla gorilla - Homo sapiens	108	41.0	41.0	1.0	1.0
Gorilla gorilla gorilla - Macaca mulatta	372	168.3	168.3	5.7	5.7
Gorilla gorilla gorilla - Meleagris gallopavo	1143	510.5	512.9	45.5	43.1
Gorilla gorilla gorilla - Monodelphis domestica	872	398.1	399.5	25.9	24.5
Gorilla gorilla gorilla - Mus musculus	553	206.9	209.2	57.6	55.3
Gorilla gorilla gorilla - Oryctolagus cuniculus	384	167.9	168.0	12.1	12.0
Gorilla gorilla gorilla - Ovis aries	596	264.8	265.1	19.7	19.4
Gorilla gorilla gorilla - Pan troglodytes	138	55.0	54.9	2.0	2.1
Gorilla gorilla gorilla - Papio anubis	275	121.9	121.9	3.6	3.6
Gorilla gorilla gorilla - Pongo abelii	165	69.0	69.0	1.5	1.5
Gorilla gorilla gorilla - Rattus norvegicus	987	419.5	423.4	62.0	58.1
Gorilla gorilla gorilla - Sus scrofa	1469	700.5	700.9	22.0	21.6
Gorilla gorilla gorilla - Taeniopygia guttata	964	431.0	432.1	34.5	33.4
Homo sapiens - Macaca mulatta	246	108.4	108.4	3.1	3.1
Homo sapiens - Meleagris gallopavo	1101	489.2	491.8	45.8	43.2
Homo sapiens - Monodelphis domestica	836	379.6	381.2	26.9	25.3
Homo sapiens - Mus musculus	493	176.9	179.4	58.1	55.6
Homo sapiens - Oryctolagus cuniculus	340	146.9	147.0	11.6	11.5
Homo sapiens - Ovis aries	507	221.4	221.6	18.6	18.4
Homo sapiens - Pan troglodytes	104	38.0	37.9	2.0	2.1
Homo sapiens - Papio anubis	204	89.0	89.0	1.5	1.5
Homo sapiens - Pongo abelii	119	46.0	46.0	1.5	1.5
Homo sapiens - Rattus norvegicus	910	381.7	385.7	61.8	57.8
Homo sapiens - Sus scrofa	1451	692.0	692.4	22.0	21.6
Homo sapiens - Taeniopygia guttata	925	412.0	413.1	34.0	32.9
Macaca mulatta - Meleagris gallopavo	1174	521.5	524.5	50.0	47.0
Macaca mulatta - Monodelphis domestica	911	414.6	416.5	30.4	28.5
Macaca mulatta - Mus musculus	573	214.6	216.9	61.4	59.1
Macaca mulatta - Oryctolagus cuniculus	407	179.8	179.9	12.7	12.6
Macaca mulatta - Ovis aries	621	276.1	276.4	20.9	20.6
Macaca mulatta - Pan troglodytes	267	115.8	115.8	5.7	5.7
Macaca mulatta - Papio anubis	114	45.5	45.5	1.0	1.0
Macaca mulatta - Pongo abelii	274	120.4	120.4	4.6	4.6
Macaca mulatta - Rattus norvegicus	1016	434.4	438.3	63.1	59.2
Macaca mulatta - Sus scrofa	1505	716.9	717.4	25.1	24.6
Macaca mulatta - Taeniopygia guttata	997	444.2	445.6	37.8	36.4
Meleagris gallopavo - Monodelphis domestica	1120	524.3	525.7	20.2	18.8
Meleagris gallopavo - Mus musculus	1179	489.9	498.6	84.1	75.4
Meleagris gallopavo - Oryctolagus cuniculus	860	381.2	382.8	33.3	31.7
Meleagris gallopavo - Ovis aries	1154	509.8	513.4	51.7	48.1
Meleagris gallopavo - Pan troglodytes	1089	487.4	489.7	41.6	39.3
Meleagris gallopavo - Papio anubis	1104	494.3	496.7	42.2	39.8

Meleagris gallopavo - Pongo abelii	1074	483.8	485.6	37.7	35.9
Meleagris gallopavo - Rattus norvegicus	1427	608.5	620.0	89.5	78.0
Meleagris gallopavo - Sus scrofa	1434	652.0	655.9	49.5	45.6
Meleagris gallopavo - Taeniopygia guttata	542	245.2	245.3	9.3	9.2
Monodelphis domestica - Mus musculus	948	376.6	391.7	87.4	72.3
Monodelphis domestica - Oryctolagus cuniculus	665	287.0	290.4	34.5	31.1
Monodelphis domestica - Ovis aries	965	420.7	426.6	48.3	42.4
Monodelphis domestica - Pan troglodytes	812	355.0	358.5	39.0	35.5
Monodelphis domestica - Papio anubis	802	352.9	356.1	37.6	34.4
Monodelphis domestica - Pongo abelii	786	347.2	349.7	33.8	31.3
Monodelphis domestica - Rattus norvegicus	1285	529.1	553.7	102.9	78.3
Monodelphis domestica - Sus scrofa	1576	731.1	737.5	47.4	41.0
Monodelphis domestica - Taeniopygia guttata	969	443.1	444.9	24.9	23.1
Mus musculus - Oryctolagus cuniculus	422	144.2	146.9	55.8	53.1
Mus musculus - Ovis aries	607	219.6	223.8	70.4	66.2
Mus musculus - Pan troglodytes	506	180.3	182.8	60.7	58.2
Mus musculus - Papio anubis	491	175.9	178.1	59.1	56.9
Mus musculus - Pongo abelii	453	155.4	157.6	59.1	56.9
Mus musculus - Rattus norvegicus	780	358.7	359.0	20.8	20.5
Mus musculus - Sus scrofa	1532	690.7	694.4	65.3	61.6
Mus musculus - Taeniopygia guttata	1010	414.1	419.8	74.4	68.7
Oryctolagus cuniculus - Ovis aries	447	193.0	193.2	17.0	16.8
Oryctolagus cuniculus - Pan troglodytes	339	147.5	147.6	10.0	9.9
Oryctolagus cuniculus - Papio anubis	327	141.9	142.0	10.6	10.5
Oryctolagus cuniculus - Pongo abelii	311	131.9	132.0	11.6	11.5
Oryctolagus cuniculus - Rattus norvegicus	730	300.8	304.3	53.2	49.7
Oryctolagus cuniculus - Sus scrofa	1072	505.2	505.6	19.8	19.4
Oryctolagus cuniculus - Taeniopygia guttata	739	325.2	326.1	27.8	26.9
Ovis aries - Pan troglodytes	521	229.4	229.7	17.6	17.3
Ovis aries - Papio anubis	507	224.5	224.7	15.5	15.3
Ovis aries - Pongo abelii	495	218.7	218.8	15.3	15.2
Ovis aries - Rattus norvegicus	1032	434.0	439.5	68.5	63.0
Ovis aries - Sus scrofa	1481	706.2	706.7	20.8	20.3
Ovis aries - Taeniopygia guttata	997	440.3	442.2	41.7	39.8
Pan troglodytes - Papio anubis	232	101.4	101.4	2.6	2.6
Pan troglodytes - Pongo abelii	147	60.5	60.5	1.0	1.0
Pan troglodytes - Rattus norvegicus	886	370.5	374.2	60.5	56.8
Pan troglodytes - Sus scrofa	1361	647.7	648.1	20.8	20.4
Pan troglodytes - Taeniopygia guttata	906	403.8	404.8	32.7	31.7
Papio anubis - Pongo abelii	175	72.4	72.4	3.1	3.1
Papio anubis - Rattus norvegicus	902	379.6	383.3	60.9	57.2
Papio anubis - Sus scrofa	1388	660.8	661.2	22.7	22.3
Papio anubis - Taeniopygia guttata	924	409.5	410.8	36.0	34.7
Pongo abelii - Rattus norvegicus	897	376.9	380.3	59.6	56.2
Pongo abelii - Sus scrofa	1377	656.9	657.2	19.6	19.3
Pongo abelii - Taeniopygia guttata	901	400.7	401.7	33.3	32.3
Rattus norvegicus - Sus scrofa	1823	834.2	839.9	66.8	61.1
Rattus norvegicus - Taeniopygia guttata	1293	550.4	558.5	79.6	71.5
Sus scrofa - Taeniopygia guttata	1311	595.3	597.8	43.7	41.2

Table 8: Estimated distances for all species combinations of phylogenetic tree Fig. 15

Distances represented in: 1) number of synteny blocks measured with PhylDiag, 2 + 4) estimated number of inversions (Inv) and reciprocal translocations (Transl) as estimated by using Mazowita et al. (2006), and 3+5) estimated number of inversions and reciprocal translocations by using a modified estimator (InvUnif and TranslUnif).

Declaration of Authorship

I hereby confirm that I have authored this Master's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, August 16, 2015

Lucas Tittmann